A MACHINE LEARNING APPROACH TO THE IDENTIFICATION OF TRANSLATIONAL LANGUAGE: AN INQUIRY INTO TRANSLATIONESE LEARNING MODELS

Iustina-Narcisa Ilisei

A thesis submitted in partial fulfilment of the
requirements of the University of Wolverhampton
for the degree of Doctor of Philosophy

12th October 2012

This work or any part thereof has not previously been presented in any form to the University or to any other body whether for the purposes of assessment, publication or for any other purpose (unless otherwise indicated). Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

The right of Iustina-Narcisa Ilisei to be identified as author of this work is asserted in accordance with ss.77 and 78 of the Copyright, Designs and Patents Act 1988. At this date copyright is owned by the author.

Signature .....
Date .....

Abstract

In the field of Descriptive Translation Studies, *translationese* refers to the specific traits that characterise the language used in translations. While translationese has been often investigated to illustrate that translational language is different from non-translational language, scholars have also proposed a set of hypotheses which may characterise such differences. In the quest for the validation of these hypotheses, embracing corpus-based techniques had a well-known impact in the domain, leading to several advances in the past twenty years. Despite extensive research, however, there are no universally recognised characteristics of translational language, nor universally recognised patterns likely to occur within translational language. This thesis addresses these issues, with a less used approach in the field of Descriptive Translation Studies, by investigating the nature of translational language from a machine learning perspective.

While the main focus is on analysing translationese, this thesis investigates two related sub-hypotheses: simplification and explicitation. To this end, a multilingual learning framework is designed and implemented for the identification of translational language. The framework is modelled as a categorisation task, the learning techniques having the major goal to automatically learn to distinguish between translated and non-translated

texts. The second and third major goals of this research are the retrieval of the recurring patterns that are revealed in the process of solving the task of categorisation, as well as the ranking of the most influential characteristics used to accomplish the learning task. These aims are fulfilled by implementing a system that adopts the machine learning methodology proposed in this research.

The learning framework proves to be an adaptable multilingual framework for the investigation of the nature of translational language, its adaptability being illustrated in this thesis by applying it to the investigation of two languages: Spanish and Romanian. In this thesis, different research scenarios and learning models are experimented with in order to assess to what extent translated texts can be differentiated from non-translated texts in certain contexts. The findings show that machine learning algorithms, aggregating a large set of potentially discriminative characteristics for translational language, are able to differentiate translated texts from non-translated ones with high scores. The evaluation experiments report performance values such as accuracy, precision, recall, and F-measure on two datasets.

The present research is situated at the confluence of three areas, more precisely: Descriptive Translation Studies, Machine Learning and Natural Language Processing, justifying the need to combine these fields for the investigation of translationese and translational hypotheses.

Acknowledgements

The present research would not have been possible without the support of several persons: my supervisors, my colleagues, my family and my friends. I am deeply grateful to all of you.

First and foremost, I would like to thank my Director of Studies, Prof. Ruslan Mitkov, for giving me the privilege of working in one of the best research groups in computational linguistics, for encouraging me to start this journey, and for his guidance throughout my doctoral studies. I am grateful to my supervisor, Prof. Gloria Corpas, for her constant encouragements, and for allowing me to work on her corpora to conduct part of my experiments for this thesis. My deepest gratitude to Prof. Diana Inkpen, my supervisor, whose prompt and constructive feedback, unbelievable patience and exemplary support were invaluable for this thesis. I am also grateful to my examiners, Mark Shuttleworth and Georgios Paltoglou, for their appreciation and insightful comments on my thesis. A big thank you to Prof. Dan Cristea, my former professor during my master's studies, whose contagious enthusiasm for computational linguistics kindled my interest in this domain.

Thank you to all my colleagues and former members of the research group for their friendship that made my Ph.D. journey more pleasant.

Special thanks to Laura Hasler, Georgiana Marşic, Luz Rello, Lucia Specia, Irina Temnikova, Andrea Varga, Iustin Dornescu, Claudiu Mihăilă, Constantin Orăsan and Viktor Pekar for many stimulating discussions that helped shape my ideas. A million thanks to Georgiana Marşic for taking the time to read, proofread, and provide constructive feedback on the entire thesis. I am also grateful to Claudiu Mihăilă, Constantin Orăsan and Irina Temnikova for their careful reading and suggestions on my key chapters. A big thank you to those who proofread distinct parts of my manuscript: Alison Carminke, Emma Franklin, Zoe Harrison and Erin Stokes.

I am deeply and forever indebted to my family: my amazing parents, my eternally enthusiastic brother and my wise grandmother. There are no sufficient words to thank them for being incredibly understanding and supportive, helping me every single step of the way on this difficult journey. To my beloved brother, who always stood by me, a massive Thank You! Special thanks to Javier, my boyfriend and my best friend, for being incredibly supportive, understanding and, most of all, patiently putting up with all the stress and the difficulties posed by conducting this research. I am also grateful to my close friends from Romania, England and across the world (you know who you are!), and especially to Elena Cernat, for always believing in me and encouraging me.

Finally, I would like to express my sincere gratitude for the financial support provided by Research Group of Computational Linguistics, University of Wolverhampton, United Kingdom at the beginning of my research journey, and by Inatech SRL, Romania in the last years of my doctoral studies.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

API	Application Programming Interface
ARFF	Attribute Relation File Format
AZP	Anaphoric Zero Pronoun
BNet	BayesNet
DTS	Descriptive Translation Studies
DTree	Decision Tree Learning Algorithm
FE	Feature Extractor
IB1	Nearest-neighbour Learning Algorithm
ML	Machine Learning
MT	Machine Translation
MTPC	Original Medical Texts Written by Professionals
MTP	Medical Translations Written by Professionals
MTS	Technical Translations Written by Professionals
MTSC	Original Medical Texts Written by Students
MTxt	Texts from the Medical Domain
NB	Naïve Bayes Learning Algorithm
NLP	Natural Language Processing
SL	Source Language
SMT	Statistical Machine Translation
SVM	Support Vector Machine Learning Algorithm
RoTC	Romanian Translational Corpus
TEC	Translational English Corpus
TL	Target Language
TS	Translation Studies
TT	Technical Translated Texts
TTC	Original Technical Texts Written by Professionals
TTxt	Texts from the Technical Domain
TU	Translation Universals
TM	Translation Memory
TMs	Translation Memory Systems
WEKA	Waikato Environment for Knowledge Analysis Machine Learning Software
URL	Uniform Resource Locator
XML	eXtended Markup Language
ZeroR	The 0-R Learning Algorithm
ZP	Zero Pronoun

Chapter 1

Introduction

1.1 Overview

Translationese, the phenomenon which hypothesises the existence of specific features that characterise translational language as opposed to non-translational[1], has been largely investigated within the Descriptive Translation Studies domain. Embracing corpus-based techniques in the quest for the validation of various translational hypotheses has had a well-known impact in this domain, leading to several advances in the last twenty years. Yet, despite extensive research, the characteristics of translational language have not been universally recognised, nor have patterns been universally recognised as likely to occur within translational language. This is because important issues need to be addressed in order to provide a methodology which can facilitate such advances in the domain.

The first issue is that the corpus-based approach often illustrates results which are difficult to interpret from any translational hypothesis

[1]More details on the concept of translationese follow in Section 2.2.2.

point of view. In addition, due to the limitations of manual analysis, most of the studies adopt small-scale resources, investigate a low number of features[2] and often conduct experiments for only one hypothesis at a time. These issues are addressed in this thesis by proposing a novel approach, able to use large amounts of data, and providing a perspective over translational language which aggregates several features at the same time.

Second, the current approach largely adopted within translation research does not allow a comparison of the findings among scholars. This occurs because of the lack of a common methodology: whereas some scholars investigate the potential features in terms of overall word frequencies, others analyse a limited set of items, often language-dependent features. However, the interpretation of the results falls into the same translational hypothesis. This issue prevents the investigation of the validity of the potentially universal features in translational language. Therefore, a common methodology based on the investigation of multilingual features is required within the domain. This thesis addresses this necessity by analysing multilingual attributes, namely features which can also be retrieved and investigated for other languages.

Third, despite extensive research into potential features of simplification or explicitation[3], no study investigates and ranks these features according to the extent to which the characteristic appears to make a considerable distinction between translated and non-translated texts. The present research bridges this gap by providing a ranking of

[2]Throughout this thesis, the terms *feature* and *characteristic* are synonymous and, thus, they are used interchangeably. More details are provided in Section 3.3.1.2.

[3]More details on the simplification and explicitation hypotheses follow in Section 2.2.3.2 and 2.2.3.3.

the most influential features in a task of categorisation between translated and non-translated texts.

Moreover, it has been recently pointed out that novel knowledge, in terms of the recurring patterns likely to occur within translational language, is necessary in order to deepen the understanding of the phenomena of translation (Chesterman, 2004*a*) and, thus, to facilitate developments towards more accurate translations. This research bridges that gap by adopting learning algorithms that are able to retrieve patterns that can distinguish between translated and non-translated texts.

Throughout this thesis, by *recurring patterns of translations* (or sometimes hereafter referred to as *translational patterns*) is understood a correlation between one or more quantifiable characteristics (also referred to as features, variables or attributes) and the translated texts[4].

Considering the shortcomings identified above, one suggestion for improvement would be to create an adaptable, multilingual methodology able to simultaneously handle a large set of features regarding the same data, and providing a ranking across the features explored. This type of methodology would enhance the perspective over the phenomena occurring in translational language.

The methodology proposed in the current research investigates translationese, focusing on two of its related sub-hypotheses: simplification and explicitation. At this point, it is important to emphasise that the 'translationese' terminology is used without any negative connotation in the present research, and it is seen as an umbrella term for all the

[4]More details and examples of such patterns are provided in Section 4.3.

characteristics specific to translated text, regardless of what other sub-hypothesis a particular characteristic may support at the same time[5].

The analysis of the translational hypotheses is conducted by employing machine learning techniques able to automatically distinguish between translated and non-translated texts. The present approach is situated at a confluence of three areas: translational hypotheses research; machine learning; and natural language processing.

To the best of the author's knowledge, the expression *translational hypotheses research* is coined first in the present thesis, and it refers to the research conducted on the well-known translation universals, tendencies or norms of translation. In this thesis, these tendencies are seen as hypotheses only, and thus, the corresponding terminology is used[6].

The three areas mentioned earlier are relevant for the following reasons: first, this research is investigating potential hypotheses of translation, which places this study principally within translation research; second, the method chosen for investigation is the machine learning approach; third, the approach adopted requires a set of characteristics which are automatically retrieved using natural language processing tools.

[5]Several remarks on the definition of translationese as well as the meanings of the term used in the literature are pointed out in Section 2.2.2.

[6]See the justification of the *translational hypothesis* concept in Section 2.2.3.

1.2 Aims and Objectives

The **main aim** of the thesis has three aspects. First, this research aims to provide the basis of a multilingual methodology through which translational hypotheses can be investigated. The approach adopted is automatically learning to categorise texts as translated or non-translated. The translational hypotheses addressed in this research are: translationese, simplification and explicitation.

The **second major aim** of this thesis is to produce a ranking of the most relevant features (aggregated in this study) which distinguish between translated and non-translated texts.

The **third main aim** of the thesis is to retrieve recurring patterns of translational language, based on which the model is able to learn to distinguish between translated and non-translated texts.

Thus, considering these major aims of this study, the **research questions** addressed by this thesis are as follows:

1. how is it possible to model a multilingual environment that makes the correlation between potential characteristics of translational language and translated texts? In other words, how is it possible to model a system that is able to point out the characteristics of translational language?

2. to what extent can translational and non-translational language be automatically distinguished?

3. to what extent can features of translational language hypothesised in the literature distinguish translated texts from non-translated ones? More precisely:

   (a) to what extent can the potential features of the simplification hypothesis differentiate translated texts from non-translated ones?

   (b) to what extent can the potential features of the explicitation hypothesis differentiate translated texts from non-translated ones?

   (c) which are the most relevant features (from those under investigation) that influence the task of categorisation between translated and non-translated texts?

4. given a set of potential features of translational language, to what extent can recurring patterns likely to occur within translational language be automatically retrieved?

   (a) given a set of potential features of simplification, to what extent are these features involved within the translational patterns?

   (b) given a set of potential features of explicitation, to what extent are these features involved within the translational patterns?

5. to what extent can a potential feature of translational language distinguish between translated and non-translated texts?

To achieve the aims outlined, and to address the research questions listed above, the following **objectives** need to be fulfilled:

Objective 1

Related research: To conduct an extensive investigation into the existing research studies and their main approaches to the investigation of translational language. Such an overview is necessary to contextualise the strengths and drawbacks of the existing research, and to determine any gaps in the existing research in order to identify further lines of investigation needed in the domain.

Objective 2

Line of research adopted: To propose a novel methodology for the investigation of translational hypotheses and justify its need within the current research context, emphasising its main strengths.

To identify the approach needed in order to design a *multilingual* and *fully automatic* methodology which is required to pave the way towards the possibility of investigating whether the proposed translational hypotheses within the domain occur universally, regardless of source or target languages.

Objective 3

Data acquisition: To select the type of corpus required for the investigation of the selected translational hypotheses, namely translationese, simplification and explicitation.

Since all these hypotheses refer to features specific of translated texts as opposed to non-translated texts, comparable corpora are required for the research. These resources are available for distinct languages. In this work, Spanish and Romanian corpora are used. Given that for

Romanian a comparable corpus of this kind did not exist, this work undertakes the task of compiling the necessary corpus.

Objective 4

Tools needed: To gather the tools necessary for the extraction of features to be investigated for the translational hypothesis selected. As the methodology implies, natural language processing tools are required in the research process for the automatic extraction of the features under investigation. As these types of resource are language-dependent, the tools need to handle the selected languages, Spanish and Romanian.

Objective 5

Methodology: To identify potential features of translational language to design the learning model framework necessary for the investigation of translational hypotheses.

This learning model is named *translationese generic learning model*.

Objective 5.1

To select the features considered by scholars within the domain to support simplification and explicitation features, respectively. To design research scenarios which assess whether simplification features, or explicitation features, respectively, contribute to the classification task between translated and non-translated texts.

Objective 5.2

To design a learning model to investigate the simplification features, model hereafter referred to as the *simplification learning model*.

Objective 5.3

Design a learning model to analyse the explicitation features, model hereafter referred to as the *explicitation learning model*.

Objective 6

To implement the learning models designed and to assess their findings.

Objective 6.1

To implement the translationese generic learning model and to analyse its findings.

Objective 6.2

To investigate to what extent the simplification features influence the translationese generic learning model. To implement the necessary learning model in order to assess to what extent the simplification features contribute to the generic learning model. More precisely, to remove the simplification features from the translationese generic learning model. The resulting learning model is named the *excluding simplification learning model*. To compare the performance obtained between the two learning models, the translationese generic learning model and the excluding simplification learning model, and to analyse the outcomes.

Objective 6.3

To investigate to what extent the explicitation features influence the translationese generic learning model. To implement the necessary learning model in order to assess the contribution

of the explicitation features to the generic learning model. More precisely, to remove the explicitation features from the translationese generic learning model. The resulting learning model is named the *excluding explicitation learning model*. To compare the performance obtained between the two learning models, the translationese generic learning model and the excluding explicitation learning model, and to analyse the outcomes.

Objective 6.4

To implement and analyse the results obtained from the learning models designed, namely, the simplification learning model and the explicitation learning model.

Objective 6.5

To analyse to what extent a learning model that relies on only one attribute at a time can handle the task of distinguishing between translated and non-translated texts.

Objective 6.6

To identify the strengths and the limitations of the proposed methodology.

Objective 7

To suggest future directions and endeavours for this area of research.

At this point, it is important to emphasise several aspects which are not the subject of the current research study:

- This is not a qualitative analysis of translational language; the present thesis assesses translational versus non-translational language in a

quantitative manner in order to design an adaptable, multilingual environment able to categorise the two types of text.

- This study does not investigate if specific tendencies of translated text occur universally, regardless of the source or target languages, and, hence, whether they occur in any translational text. It proposes a multilingual model, easily adaptable for further investigations on other languages.

- This study does not assess whether translationese is a desirable or non-desirable phenomenon which occurs in texts; translationese is understood only as the hypothesis which indicates that translational language has specific features in comparison to non-translational language. The present research uses the translationese terminology without any negative connotation, it only focuses on a computational approach for investigating this translational hypothesis.

- This study does not investigate the process of translation: it analyses the product of translation.

- This study does not analyse whether the distinctive features which appear to influence the learning model are due to conscious or unconscious tendencies of professional translators.

1.3 Original Contributions

To summarise, the main **original contributions** of this thesis are as follows:

1. a novel multilingual methodology to investigate translational hypotheses;
   (a) a novel approach to assess the potential features of the simplification hypothesis;
   (b) a novel approach to assess the potential features of the explicitation hypothesis;
2. novel knowledge regarding recurring translational patterns (being the first computational study which addresses this aim)[7];
3. the first study which provides a ranking among the features able to categorise translational and non-translational language;
4. a novel resource for the translation studies domain to be created as a by-product of this research: the comparable corpus for Romanian.

To achieve this, an extensive review of the current research studies and approaches to the investigation of the nature of translational language has been undertaken. The strengths and limitations of the research context have been assessed, and some research gaps and issues within the domain

[7]Note that whereas the simplification and explicitation phenomena are seen as potential hypotheses of translational language throughout this thesis, translational patterns are seen as a correlation between one or more quantifiable characteristics and the nature of translational language. Details on such patterns are provided in Section 4.2.

are addressed by this research. This thesis analyses potential features of translationese in general, focusing on the simplification and explicitation hypotheses. All three hypotheses are seen in this thesis as translational hypotheses.

The **first contribution**, the novel multilingual methodology, arises at the confluence of three disciplines: translation studies; machine learning; and natural language processing. The theoretical concepts come from translation studies, and the methodology and tools pertain to machine learning and natural language processing domains. The need to combine these three areas is justified as follows: to extract the potential features for a certain translational hypothesis, tools from natural language processing are employed. To assess whether a feature, or a set of combined features, are in fact specific to translational language, a task of distinguishing translated from non-translated texts based on these features is proposed using the machine learning approach[8].

Since the hypotheses assume that these features are specific to translational language, then translated texts can presumably be identified based on them. The task of categorisation is automated and, moreover, statistical algorithms are employed in order to *learn* to categorise between these two types of text. In this way, a machine learning approach is adopted for the investigation of translational hypotheses.

The multilingual aspect of the methodology is assured by the use of multilingual attributes within the learning framework. In addition, the

[8]The potential interest that the machine learning techniques and natural language processing tools might hold for the translation studies domain is presented in Section 4.2.2 and 4.2.3.

thesis conducts experiments on two languages, Spanish and Romanian, emphasising the flexible character of the methodology proposed.

The **second contribution** is achieved by employing learning algorithms which identify recurring patterns based on which they learn to distinguish between translated and non-translated texts. Furthermore, the **third contribution** of this research, the ranking of the characteristics considered in the learning model, is obtained by employing specific ranking filters largely used within the machine learning domain: Information Gain and Chi-squared ranking filters. These algorithms retrieve the most relevant characteristics used in the categorisation task between translated and non-translated texts.

The **fourth contribution** represents the resource compiled for these experiments: the Romanian translational corpus, RoTC. The scarcity of Romanian resources is overcome by the compilation of a new corpus, assembled according to the needs of this research. The compilation process is detailed in Chapter 3.

1.4 Applications of this Research

Besides the contributions of this research within Descriptive Translation Studies, the contributions of the present learning model for distinct applications are outlined below. The main contribution is that the ability to identify translated texts can improve the performance of natural language processing (NLP) applications, such as: machine translation, automatic compilation of parallel corpora, and cross-lingual plagiarism detection.

Also, the model proposed in this thesis can be integrated within further applications for training students of translation. The following paragraphs detail how this can be achieved.

Research studies within the natural language processing field, corroborating the present research, have pointed out that by creating statistical language models[9] based on translated texts, rather than on non-translated ones, the overall performance of their statistical machine translation (SMT) system improves (Koppel and Ordan, 2011; Lembersky et al., 2011; Volansky et al., 2011; Lembersky et al., 2012).

Moreover, for SMT frameworks based on large-scale web content data, it was pointed out that, by processing web language content, the framework becomes "polluted" by the errors from the translated data encountered. The term *translated data* on the web refers to texts either translated by humans (usually untrained in translation), or by automatic machine translation tools. The errors create statistically relevant predictors and, consequently, the performance of the statistical machine translation system deteriorates. Therefore, the framework needs to extract only the non-translated texts from the website data and disregard the translated data. To be able to filter out translated texts, the present research proposes a model which can be adapted and integrated within their framework[10].

Another application that can benefit from the identification of translated texts is the automatic compilation of parallel corpora (Resnik

[9]Statistical language models are designed to assign probabilities to string of words or tokens (Brants et al., 2007), and are used to improve the fluency of generated translated text. See the formal definition in Brants et al. (2007, p. 858).

[10]The article named "*An 'economic burden' Google can no longer bear?*", published in the Technology section of the 'The Atlantic' news portal, can be found in Appendix D. Website last accessed on 18th August 2012: `http://www.theatlantic.com/technology/archive/2011/06/an-economic-burden-google-can-no-longer-bear/240283/`

and Smith, 2003). The current research can improve the methodology used for web-based parallel corpus extractors by retrieving the candidate parallel texts. A distinct natural language processing branch which can integrate the module of automatic identification of translated texts within its framework is the cross-lingual plagiarism detection field: when a suspected paragraph may be plagiarised from a different language (Barrón-Cedeño et al., 2010, p. 771). The present research provides the methodology to assess whether or not a suspicious text is a translation from another language.

1.5 Structure of the Thesis

This thesis comprises six chapters, in which the objectives of the research are followed systematically.

Chapter 2 provides an overview on translationese and various translational hypotheses, focusing on the relevant hypotheses to this work: simplification and explicitation. The chapter is divided into three parts: the first part presents the theoretical background, the second surveys the main approaches used in the domain, whilst the third part outlines the proposed direction of research undertaken in this thesis.

Chapter 3 provides the necessary information regarding the resources and tools required in this investigation. As this research conducts experiments on two languages, two comparable corpora are employed: for Spanish and for Romanian, both comprising translated and non-translated texts. These resources are reported in this chapter. For Romanian, the

required type of corpus was not available. Therefore, a translational comparable corpus was compiled, RoTC, according to the needs of this research. In the second part of the chapter, the machine learning tool, Weka, along with the basic concepts of the discipline are presented.

Chapter 4 reports on the methodology adopted in this thesis and justifies its need within the domain. The position of this research is discussed, situated as it is at the confluence of the three areas: translation research; machine learning; and natural language processing. However, the major goal of this research, the investigation of the nature of translational language, places this work principally in the domain of translation studies. The approach chosen for the investigation is machine learning, which requires a set of features in order to build a model. To this end, the features are extracted using natural language processing tools, thus, involving a third domain.

The machine learning approach examines the nature of translational language by modelling a task of categorising between translated and non-translated texts. The main rationale of this research is as follows: if the learning system is able to distinguish between translated and non-translated texts, then the translationese hypothesis has a strong argument in its favour.

A machine learning framework is designed for Spanish and Romanian experiments, largely having the same type of characteristic. Their differences are reported and justified. The experiments are categorised in five research scenarios, presented within the chapter. The learning models proposed for the investigation of translationese, simplification and

explicitation are reported and explained: the translationese generic learning model, the simplification learning model and the explicitation learning model.

Chapter 5 reports the results obtained for both languages, listed in the order that the experiments were conducted. The chapter is divided into two sections: the Spanish and the Romanian experiments. Each research scenario is reported and the findings are analysed mainly from the perspective of translationese, as well as from the simplification and explicitation viewpoints. Finally, the results of the overall framework are discussed for each language, providing a comparison to related work. The chapter finishes by highlighting the strengths and the shortcomings of the current methodology.

In **Chapter 6**, the concluding remarks of this research are reported. The chapter revisits the aims and the objectives of the thesis, discussing to what extent they have been accomplished in the experiments conducted. The thesis finishes by suggesting future avenues of research.

Chapter 2

Related Research

2.1 Overview

This chapter describes the notion of translationese in the context of the wider domain of Translation Studies, and presents previous work on the translational hypotheses relevant to the experiments reported in the present thesis. The objective is to discuss the most important studies relevant to this research, and not to exhaustively describe all the investigations reported on translational hypotheses currently available within the domain.

The chapter is divided in three sections: first, the theoretical background is introduced focusing on the hypotheses addressed by this thesis: translationese, simplification and explicitation hypotheses. In Section 2.3, the main approaches used in the investigation of these translational hypotheses, as well as their well-known findings, are outlined. Most of the existing studies prefer a corpus-based approach, presented in Section 2.3.1, whereas an insignificant number of studies adopt the use

of machine learning techniques in their analysis, research highlighted in Section 2.3.2.

The main strengths and shortcomings of the existing research on the translational hypotheses are emphasised in Section 2.3.3, whereas in Section 2.4, a potential line of research for the advancement of the domain is suggested.

2.2 Theoretical Background

In this section, the controversial notion of translationese is located within its own domain, namely Descriptive Translation Studies. First, translation and the discipline of translation studies are outlined, and in the second part of this section, the translationese phenomenon is highlighted along with related translational hypotheses.

2.2.1 Introduction to Translation and Translation Studies

Although the notion of translation has been used with slightly different meanings over time, it can be described as the mechanism of "transferring a written text from source language to target language, conducted by a translator, in a specific socio-cultural context"(Hatim and Munday, 2004, p. 6).

Because of the different nuances in its meaning, translation is categorised into the following three types: *intralingual* translation (or

'rewording'), *interlingual* translation (known as the translation proper), and *intersemiotic* translation (e.g., when a written text is being translated into music, film or painting) (Jakobson, 1959/2000). The most thoroughly investigated category so far is interlingual translation, a term which can refer to:

- the general subject field;
- a text that has been translated (the final product);
- the act of producing the translation (Munday, 2008).

The focus of the present research is on the final product, the newly-produced, translated text, and the term *translation* is used in this sense throughout the thesis.

The main problem of translation arises as a consequence of a natural, expected misalignment between the source and target languages, which prevents ideal, exact translations. This occurs because no two languages have the same set of concepts and cultural backgrounds and, thus, the translation product cannot be entirely mapped onto the target language. This is a well-known fact in the literature, highlighted and explained by Nida (2000): as no two languages are identical, "it stands to reason that there can be no absolute correspondence between languages. Hence, there can be no fully exact translations"(Nida, 2000, p. 126). A degree of interpretation on the translator's side prevails, regardless of how much translators endeavour to provide the best translation possible.

Given that the main purpose of translation is to enable a cross-cultural communication event, both language and culture are thus involved in this

process. Because of a great variation in terms of beliefs, ways of thinking, and values across distinct nations, the translation process has a potential inherent ability of creating misunderstandings and/or misinterpretations as the message is carried over into the target language. The complexity of translation is pointed out by several studies, its nature being investigated from a wide range of viewpoints and disciplines.

The domain has evolved over the last sixty years and the academic discipline which provides the theoretical and methodological framework for translation was named *Translation Studies*, as Holmes (1988) suggested. In the fifties and sixties, more and more linguistic approaches were adopted in translation research (Vinay, 1958; Nida, 1964), whilst in the eighties, a distinction emerged between the theoretical translation studies and the teaching side of this discipline (i.e., the "technique in foreign-language instruction" category and "translator training"(Holmes, 1988), respectively).

Attempting to structure the newly formed academic discipline, Holmes (1988, p. 71) proposes a conceptual map, and in his classification a main distinction is made between the 'pure' and the 'applied' sides of this domain. The objectives of the 'pure' side are to describe the translation process and translated text as they reveal themselves, as they are perceived, and to propose general laws on the basis of which this phenomenon can be explained and predicted. For the latter category, the applied side of the domain, the following main goals are outlined: the training process of professional translators (such as teaching methods or testing techniques), the aids necessary for translation (dictionaries, grammars,

computer-assisted translation), and translation criticism (the evaluation stage of a translated text, reviews of published translations).

Note that the study of translation is a multilingual and interdisciplinary topic by its own nature: it is multilingual because it comprises any variety of pairs of languages, and interdisciplinary because the field benefits from diverse knowledge coming from a wide range of disciplines (Munday, 2008). Figure 2.1 illustrates the disciplines which interact with the translation phenomenon (Hatim and Munday, 2004, p. 8).

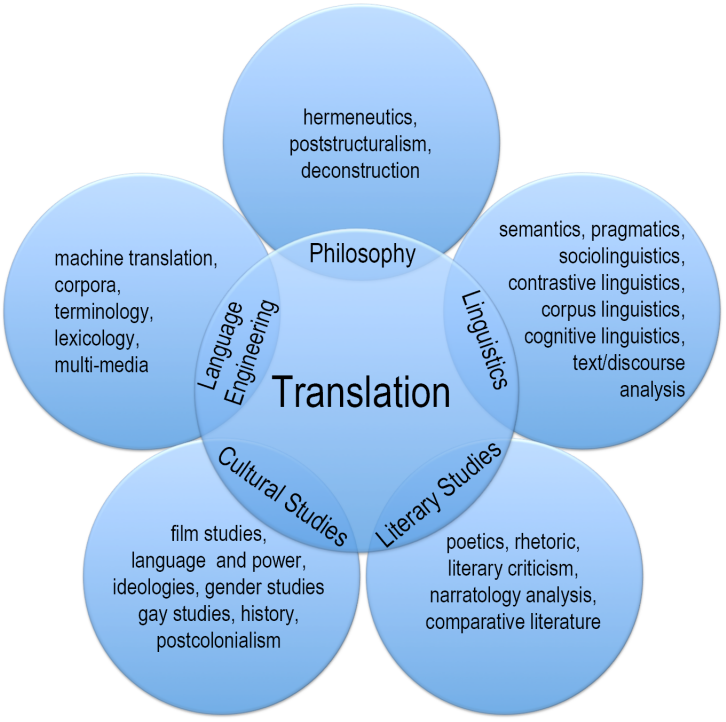


Figure 2.1: Disciplines Interfacing with Translation Studies (Hatim and Munday, 2004, p. 8)

The influences from the other disciplines have varied through time (Munday, 2008, pp. 13-17). In the sixties, a large number of studies on translation adopted the perspective of contrastive linguistics, whilst

more recent research relates to cultural studies or computing and media, such as audiovisual translation (Diaz Cintas and Remael, 2007; Chaume, 2004), localisation and globalisation (Cronin, 2003; Pym, 2004), translation technology with its machine translation and translation memory frameworks, and corpus-based translation studies (Olohan, 2004; Corpas, 2008).

As a consequence of its interdisciplinary character, the discipline of translation studies allows for the application of distinct research methods, otherwise specific to other fields. However, the downside is that the domain lacks its own theoretical and methodological basis, being still in the process of incorporating theories and research methods from the disciplines it mostly relates to, such as linguistics, corpus linguistics, or cultural studies (Williams and Chesterman, 2002; Tymoczko, 2005).

Considering the academic fields illustrated in Figure 2.1, the present research can be included in the Language Engineering category because it is a computational approach investigating the nature of translational language. To be more precise, this research is situated at a confluence of three areas: Translation Studies and two sub-areas of Artificial Intelligence: Natural Language Processing, and Machine Learning. The adoption of a machine learning approach within the translation studies domain is strongly emphasised in this thesis because it presents a remarkable potential in the investigation of the nature of translational language and in the retrieval of translational patterns. As this approach is scarcely used in the literature, the fundamental notions are introduced within Chapter 3, and the motivation for such a turn in the domain is emphasised in Chapter 4.

In the last two decades, a shift has been noticed in the objectives of translation research, as the discipline has focused on the ways in which translational language tends to be different from the language used in non-translated texts. Scholars have hypothesised that translated language exhibits its own specific characteristics, a phenomenon covered under the term *translationese* (Gellerstam, 1986), and a set of hypotheses are proposed regarding these potential types of characteristic, hereafter referred to as *translational hypotheses*.

The fundamental notions regarding the translational hypotheses relevant for this research, namely translationese, simplification and explicitation, are further described in the next section.

2.2.2 The Translationese Phenomenon

The discipline of descriptive translation studies draws scholars' attention to various aspects of translation, attempting to describe general, universal characteristics of the nature of translation.

Scholars investigate the impact of the source language on the translations, as well as the impact on the target language, and take into account the following factors: the specific pair of languages involved; the manner in which the mapping of the message is mediated; the effects on the final product, namely, the translated text. Investigations have two distinct starting points: first, the perspective offered by the research on the potential effects caused by the source language; second, the perspective provided by the common traits hypothesised to prevail in the target text, regardless of the source language involved.

Manifestations such as translationese, the so-called translation universals, the proposed norms and laws of translated texts, seemed to capture more and more of the research community's interest at the beginning of the nineties and a considerable amount of studies focused on these descriptive hypotheses.

Since translationese was pointed out in the literature, the definition of the term has slightly varied over time. One of the first scholars points it out as *the unusual character of the language used in translations* (Toury, 1979). The author sees translationese as a linguistic system, called interlanguage[1], which "enjoys an intermediate status between source language and target language", and exhibits an "interference of these two codes [source language, target language] in the performance of the learner"(Toury, 1979, p. 223). The author also argues that translational language can be seen as a dialect (Toury, 1979, p. 228), emphasising that the target texts differ "dialectically" from the original source texts (Toury, 1980, p. 42).

Referring to the same phenomenon, Duff (1981) adopts the term *the third language* pointing out the "tyranny"over translated texts caused by the influence of the source language (Duff, 1981, p. 113). The manifestations are seen as a mixture of styles and languages comprising elements from both the source and target languages involved.

The notion of translationese has also been explained by an *analogy to motherese*, the concept from language acquisition domain which points out that parents talk differently to children than to adults. Beretta (1982) suggests that translationese is a type of simplified language, a dedicated

[1]Interlanguage is a concept which belongs to the domain of second language learning.

"foreign talk" type of language for the target readers, just like "mother talk" type of language is for children (Beretta, 1982, pp. 248-249).

A few years later, Frawley (1984) reported his view of translationese considering it a distinct sub-language in its own right which gathers influences from both source and target language, coining the term *'the third code'* in the literature:

> "translation [...] is essentially a third code which arises out of bilateral consideration of the matrix and target codes [...] it emerges as a code in its own right, setting its own standards and structural presuppositions and entailments."(Frawley, 1984, pp. 168-169)

In other words, it is stated that translational language takes over features from the source text or influenced by the source language to be able to transmit the intended message in the target language. The phenomenon is also sustained by other scholars who also argue that a target text is a "result of the confrontation of the source and target codes"(Baker, 1993, p. 245).

These concepts of translation as a type of sub-language or of the third code has a long tradition in the translation domain in terms of both foreignisation and negative evaluation (Øverås, 1998, p. 3). For several researchers the use of the term *translationese* indicates an undesirable effect within the translational language, and implicitly denotes a negative aspect of these manifestations in translated texts (Baker, 1993). The connotation is assumed because of the awkward, unusual traits which appear to create a distortion of the expected flow and the naturalness of the language. Baker

presents translationese as the phenomenon when the "unusual distribution of features is clearly a result of the translator's inexperience or lack of competence in the target language"(Baker, 1993, p. 248). Probably the reason for the terminology used and its negative connotation can be justified by the existence of a traditional prescriptive idea in the literature, according to which translations should not read as translations (Pym, 2005).

In contrast, Gellerstam (1986, p. 88) highlights that the phenomenon of translationese is not an effect that appears simply due to poor translations, is a phenomenon which reports systematic influences from source language. Nowadays the bad connotation associated with translationese begins to dissipate, the phenomenon being seen in more neutral terms (Meldrum, 2009). To note that the present research is also adopting this neutral view on translationese and its related hypotheses.

In the nineties, translationese is explained as a "hybrid language" that has a set of traits assumed to be inherited in the translation process from the source language and noticeable in the target language (Trosborg, 1997). However, these features are not seen as deviations from the proper structure of the language, but rather as unusual types, deviations from the expected norm of usage. This perspective is the foundation of a distinct translational hypothesis, called the *law of interference* (Toury, 1995).

The phenomenon is also defined as the set of linguistic features of translated texts, characteristics covered under the term "unmistakable fingerprints"(Gellerstam, 1996, p. 54), associated with the source language interference from a positive standpoint, referring to the unusual occurrences for the target language norm of usage.

More recently, it has been pointed out that translations exhibit their own specific lexico-grammatical and syntactic characteristics (Borin and Prütz, 2001; Hansen, 2003; Teich, 2003), manifestations referred to as translationese. This view over the phenomenon is also adopted in the present research.

Irrespective of the slight differences in meaning, it appears to be a consensus regarding the unavoidable occurrence of translationese in translated texts. Toury (1979, pp. 224-225) states that translationese cannot be avoided in translational language, regardless of the translator's experience, as this phenomenon, being a form of interlanguage, is a consequence of the two languages in contact. Other scholars also support and reinforce this statement (Baker, 1993; Gellerstam, 1996; McEnery and Xiao, 2002).

Regarding the directions of research on this topic, Toury (1979) has an undeniable role: he suggests that translationese should be formally analysed in "a systematic descriptive study of translation"(Toury, 1979, pp. 224-225). Moreover, he provides the fundamental basis for the development of further translational hypotheses, which began to emerge in the nineties, by pointing out the perspectives from which translationese could be viewed (Toury, 1979, p. 228):

- descriptive and contrastive linguistics, for the investigation of deviations from source and target language;
- psycholinguistics, for the observation of mental mechanisms produced at the linguistic contact;

- sociolinguistics, for the sociocultural and sociolinguistic aspects involved in the process of translation.

Another remarkable role in the development of the translation research belongs to Baker (1993). She expands the objectives of the study and invites further investigations on the general characteristics of translated texts, which all translations share irrespective of their source language. She points out the need for the development of tools that are able to identify the potential *universal features of translation*, regardless of the interference of specific linguistic systems (Baker, 1993, p. 243):

> "it will be necessary to develop tools that will enable us to identify universal features of translation, that is features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems ". (Baker, 1993, p. 243)

Baker (1993, 1996) provides a description of what a universal feature of translation may be, and one of the key terms appears to be the "typical" occurrences suspected to appear in translations. She emphasises the typical, universal character for some suggested, potential features of translational language, and furthermore, she introduces a few hypotheses in this direction known as *translation universals* (Baker, 1993, 1996). Other scholars also have made important contributions regarding the nature of translational language and different trends, norms, hypotheses or laws have been proposed (Toury, 1995; Kenny, 1998; Mauranen, 2008; Gaspari and Bernardini, 2010).

Moreover, a classification of translationese effects is suggested within the literature. According to Balaskó (2008, p. 61), translationese has two components: one consists of the features that behave differently from what is typical of the target language, whilst the other contains the features sustained by the known translation universals: simplification, explicitation, etc.

At this point, it is important to note that throughout this thesis, translationese is used as an umbrella term for all the features specific to translated text, regardless of what hypothesis a particular feature may support. A feature[2] F can be seen as a potential indicator of translationese, and at the same time F can also be accepted as an indicator which stands for any other translational hypothesis in the literature.

The next subsection provides more details about the translational hypotheses distinguished within the domain.

2.2.3 Hypotheses on Translational Language

A brief consideration of the very notion 'hypothesis' is provided. A hypothesis, according to its definition, is a tentative statement about something that "might be true or worth considering" and is categorised into three types of hypothesis: explanatory, descriptive, and interpretative (Chesterman, 2004*a*, pp. 1–2). The first category points out an explanation about a phenomenon, providing a set of probable causes and influences. Descriptive hypotheses suggest something about the features or structure

[2]The definition for feature and a few examples are provided in Section 3.3.1.2.

of a particular phenomenon, whilst the third category emphasises how that phenomenon should be understood, what it actually means.

Given the above categorisation, translational hypotheses are *descriptive*, since their fundamental goal is to assess what features are shared by all translations, irrespective of the language pairs involved. Throughout the thesis, the phrase *translational hypotheses* is preferred because it stresses the fact that the validity of these statements is still under question. It is used as an umbrella term for all the hypotheses regarding the nature of translational language: i.e., translation universals, norms, tendencies or laws of translation.

A fundamental characteristic of any hypothesis is the fact that it is based on a set of assumptions. In the case of translational hypotheses, the main underlying assumption is that any translation, regardless of the source or target language, shares a set of characteristics with other translations.

In an attempt to find these features, a set of perspectives are employed: the perspective of linguistics, which investigates the linguistic features of translational language, and the perspective of corpus linguistics, which is a suitable approach to manipulate large amounts of text in the pursuit of potential patterns and similarities. Largely, these are the main influences on the research into translational hypotheses.

Within the field of descriptive translation studies, whose goal is to analyse and describe the nature of translational language, the area of translational hypotheses is enriched by narrowing down the potential shared features of translation. As a result, besides the fundamental assumption that translations are distinct from non-translations, more assumptions are

made in order to find what it is that actually differentiates translations from originals. This proves to be a complex task and highly debated topic in the literature. More translational hypotheses appear and a well-known categorisation for these hypotheses is outlined in the next section.

2.2.3.1 Classification of Translational Hypotheses

Translational hypotheses can be divided into S-Universals and T-Universals, as pointed out by Chesterman (2004*a*): S-Universals are the hypotheses which involve a comparison between source texts and their translations, and T-Universals are the hypotheses which involve a comparison between translated texts and non-translated texts. S-universals focus on the source texts and how translators process them, whilst the latter category focuses on the target texts, i.e. the translated texts.

The first category, the S-Universals, includes the following hypotheses:

- Lengthening: translations tend to be longer than their source texts (Vinay, 1958);
- Explicitation: translations tend to be more explicit (Blum-Kulka, 1986; Klaudy, 1996; Øverås, 1998);
- Sanitisation: translations tend to have more conventional collocations (Kenny, 1998);
- Retranslation: to some extent, retranslations[3] appear to be more source culture orientated than the first translations;

[3]A retranslation is a new translation of an earlier translated text. It is typical of older, classic texts.

- Reduction of repetition (Baker, 1993);
- Reduction of complex narrative voices (Taivalkoski, 2002);
- Interference: translations are influenced by their source texts (Toury, 1995, p. 275);
- Standardisation: translations seem to be similar to each other (Toury, 1995);
- Normalisation: translations present "a tendency to exaggerate features of the target language and to conform to its typical patterns." (Baker, 1996, p. 183).

The second category, that of T-Universals, comprises the following hypotheses:

- Simplification: translations seem to be simpler than non-translations (Baker, 1996; Laviosa-Braithwaite, 1996);
- Untypical and unstable lexical patterning (Mauranen, 2000) - both in terms of frequencies of lexical items and untypical lexical patterning. This hypothesis is outlined as follows: "lexical patterning which differs from that which is found in original target language texts might be a universal feature in the language of translations"(Mauranen, 2000, p. 136);
- Under-representation hypothesis, also known as the "unique items hypothesis"[4]: fewer unique items occur in translations (Mauranen, 2008, pp. 41-42).

[4]In the literature, unique items are the terms from the target language which do not have equivalents in the source language, and hence are 'unique' in the target language.

Even though this thesis is mainly concerned with translationese and two important related hypotheses, simplification and explicitation, it is worth discussing a few well-known translational hypotheses within the domain.

Besides the important influence which Baker (1996) has had in the research on the nature of translational language, other scholars have also pointed out other remarkable aspects of this linguistic system. Toury (1995) proposes two potential '*laws of translation*': the law of growing standardisation, and the law of interference.

The *law of standardisation* states that in translational language "textual relations obtaining in the original are often modified, sometimes to the point of being totally ignored, in favour of habitual options offered by a target repertoire" (Toury, 1995, p. 268). A loss of style variation in favour of a general standardisation is thus implied by this hypothesis.

The second hypothesis proposed, named the *law of interference* (Toury, 1995, pp. 274-279), points out that the linguistic features from the source text are copied into the target text, mainly in terms of lexical and syntactic patterns. This influence on target text can be seen as both '*negative*' and '*positive*'. The first appears because of the non-typical patterns created, and the latter is perceived when the influence of the source text results in normal patterns in the target text, but the translator tends to have a preference for such patterns. However, the need to keep the distinction between negative (interference) and positive (transfer) is not necessarily felt, as Eskola (2002) argues that interference could be represented in neutral terms.

Convergence, also known as 'levelling out', is proposed by Baker and it states that translational language has the tendency "to gravitate towards the centre of a continuum"(Baker, 1996, p. 184). In the same line of thought, Laviosa (2002, p. 72) sees convergence as referring to the high level of homogeneity in translational language, i.e., "the relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features".

Dorothy Kenny, another prominent figure in translation research, focused on semantic prosody in translations, attempting to prove that translated texts are 'sanitised' versions of the source texts. As a consequence, she proposes the hypothesis of '*sanitisation*', which states that translations present a reduced connotational meaning as opposed to original texts (Kenny, 1998, p. 515).

In addition to these tendencies, there are more recent hypotheses related to the nature of translated texts: it is hypothesised that the language used in translations is similar to the language used by the learners of a foreign language, giving birth to the notion of *mediation universal* (Gaspari and Bernardini, 2010). In the same line of classification as presented above, this thesis suggests that this class of hypotheses can thus be categorised as the class of M-universals, since the pair of texts being compared is slightly different: the comparison is performed between translated texts written in language A and texts produced by second language learners, written in the same language A.

The next section discusses the simplification and explicitation hypotheses.

2.2.3.2 Simplification Hypothesis

The well-known simplification is largely formalised by Baker as "*the idea that translators subconsciously simplify language or message or both*"(Baker, 1996, p. 176). Moreover, according to Blum-Kulka and Levenston (1983); Laviosa-Braithwaite (1997), translational language is assumed to be simpler than the native language, at a lexical, syntactic and/or stylistic level.

When defining the hypothesis, Baker (1996, p. 183) suggests what types of findings would constitute evidence for this hypothesis. She suggests investigating whether translated texts have a narrower lexical range compared to original, non-translated texts in the same language. In addition, the author interprets this type of evidence "as a subconscious strategy of simplification on the part of the translators"(Baker, 1996, p. 183).

Even though simplification, as a preference for a simpler manner of transmitting the message to the target reader, is a prescriptive strategy in the area of second language acquisition, this concept is investigated from a descriptive standpoint within the translation domain (Wen, 2009, p. 34).

Before providing further details, it seems appropriate to define the very notion of simplicity. Given that the hypothesis describes the nature of a text, according to the Oxford English Dictionary (Simpson and Weiner, 1989), *'simple'* would entail that a text is:

- easily understood, presenting no difficulty to be understood;

- plain, basic, uncomplicated in form, nature or design (Simpson and Weiner, 1989).

Starting from the fundamental definition of simplicity, the following example attempts to capture the gist of the simplification hypothesis in the following pairs of non-translated and translated sentences. An instance of simplification is the preference for breaking long embedded sentences into simpler ones, as in the following examples[5] quoted from Wen (2009, p. 38):

1. Text: The jury also commented on the Fulton court, *which has been under fire for its practices in the appointment of appraisers*, guardians and administers.
   Simplified text: The jury also commented on the Fulton court. *The Fulton court has been under fire for its practices in the appointment of appraisers, guardians and administers.*

2. Text: *Needing money to pay my debts*, I forced myself to ask my friends.
   Simplified text: *I needed money to pay my debts. I forced myself to ask my friends.*

Embedded phrases pose a risk of misunderstanding and ambiguity for the reader; hence the translator would most likely avoid these situations, as the translator is supposed to facilitate correct understanding for the target audience.

[5]Note that the reference does not provide information on the source or target languages for these examples, only the simplification phenomenon being emphasised.

Next, a classification for the simplification hypothesis is observed within the literature. It has been suggested that simplification can be found at lexical, syntactic or stylistic level (Blum-Kulka and Levenston, 1983; Laviosa, 1998). For all the categories, various features that indicate the simplicity degree of a translated text are investigated in the literature, and these are further pointed out in Section 2.3.

2.2.3.3 Explicitation Hypothesis

The other translational hypothesis under investigation in the present research is explicitation. The hypothesis is defined as "the phenomenon which frequently leads to TT[6] stating source text information in a more explicit form than the original"(Shuttleworth and Cowie, 1997, p. 55). In contrast to the controversial views regarding the simplification hypothesis, the explicit characteristic of translations is fairly well established among scholars (Kamenická, 2007, p. 46).

Probably one of the first observations in the direction of this hypothesis appears in a comparative study examining French and English texts, in which explicitation is presented as:

> "the process of introducing information into the target language which is present only implicitly in the source language, but which can be derived from the context or the situation"(Vinay, 1958).

[6]In this quotation, TT is a well-known acronym in translation studies domain for target text.

Formally, explicitation is defined by Blum-Kulka (1986) as follows: "explicitation is a universal strategy inherent in the process of language mediation"(Blum-Kulka, 1986, p. 21). She suggests that the target text might become more redundant than the source text, as a consequence of the process of translation, and that this redundancy "can be" represented by a high level of "cohesive explicitness"(Blum-Kulka, 1986, p. 19). Note that there is room left for other indicators of explicitation.

The hypothesis created great interest in the research community right from the beginning. In addition, Baker (1996) reinforces explicitation by including it among the well-known translation universals she introduced in the domain. She formally defines it as the tendency to "*spell things out rather than leave them implicit*"(Baker, 1996, p. 180).

Since then a plethora of publications have discussed and investigated this hypothesis, all presumably referring to the same phenomenon. It is pointed out, however, that the term itself varies across the research community, and that the relation between explicitation/implicitation, specification/generalisation, and addition/omission needs to be clarified in these investigations (Kamenická, 2007). Along these lines, Perego (2003) points out that Nida (1964) views explicitation as one of the techniques of addition, whereas Øverås (1998) considers addition as one of the strategies of explicitation, and not the definition itself. However, most researchers leave this relation unresolved.

Fairly recently, Klaudy and Karoli (2005) refined the explicitation hypothesis by identifying it in relation to implicitation[7] and translation directionality. Moreover, a typology of explicitation is emphasised, and

[7]The process of embedding implicit, hidden information in a text.

according to Pym (2005), and Klaudy and Karoli (2005) explicitation is divided into two categories: *obligatory* (ex.1) and *voluntary* (ex.2).

The first obligatory explicitation, is required by the different language systems involved in translation, and the explicitation process in one direction is always matched by the implicitation process in the other direction. In contrast the second, voluntary explicitation, appears when the relation between explicitation and implicitation is asymmetric, and no linguistic system requires the additional information in the target text. The latter type of explicitation is considered a feature of the translation process itself (Klaudy and Karoli, 2005; Pym, 2005), a controversial observation within the literature (Becher, 2011*a*).

The authors note that optional explicitation is consistently more frequent than implicitation, and they provide examples of both. They include specification/generalisation under the phenomenon of explicitation/implicitation by connecting specification with explicitation and generalisation with implicitation. This association is argued for by some scholars, such as Kamenická (2007), and an agreement on the optional, asymmetric explicitation has not yet been reached in the literature.

In the next paragraphs, examples of each category are provided. Obligatory explicitation appears when the target language forces translators to add information not present in the source text, due to language restrictions (ex.1), whilst voluntary explicitation occurs only if translators deliberately avoid any possible misinterpretations in the texts they produce (ex.2).

1. *Source (English):* Frances liked her doctor.

   Translation (Portuguese): Frances gostava dessa médica.

   Back translation (English): Frances liked this *[female]* doctor.

2. *Source (English):* Você também gosta dela?

   Translation (Portuguese): So you like her too?

   Back translation (English): You like her too?

A well-known case of voluntary explicitation is highlighted in the following example, which is a title taken from a German article (Pym, 2005):

- *Source (German):* "Selbstverständlich besteht ein gewisses Interesse für Finnland, aber..."

- *Possible Translations (English)*[8]:

  1. Of course there is a certain interest for Finland, but...
  2. Naturally there is a certain interest for Finland, but...
  3. Obviously there still exists a certain interest in Finland, but...
  4. Of course here in Finland there exists a certain interest, but...
  5. It is self-explanatory that a certain interest for Finland is still standing, but...
  6. Of course here there exists a certain interest for Finland, but...

[8]Note that these potential translations for the context are written by Pym (2005), example encountered when he was translating the publication written by Kujamäki (2006, p. 50).

"Interest in Finland" would be more acceptable as a translation from the linguistic point of view, but the problem would be the high likelihood of misleading the reader to believe that the interest itself is located in Finland, when actually the whole article discusses interests located in Germany. It would thus be a classical case of miscommunication in translation based on insufficient information given to the target audience. For this reason, the translator prefers to restrict the ambiguity caused by the preposition "*in*" and explicitly translate as: "Of course Germans have a certain interest in Finland, but..."(Kujamäki, 2006).

Note that in translation number 5, the translator uses the phrase "it is self-explanatory that...", presumably as a way to avoid, as much as possible, any liability or responsibility that may be ascribed to the translator. It is a way to emphasise that the translator is not responsible for the message conveyed. This would account for the risk-management model proposed by Pym (2005), where translators manage the risk of non-cooperation in communication.

For Romanian, an optional explicitation is pointed out in the following example quoted from Balázs (2011, p. 310):

- *Source (English)*: "I have just returned from a visit to my landlord the *solitary* neighbour that I shall be troubled with".

- *Translation (Romanian)*: "Adineauri m-am întors din vizita făcută *posaculuí* meu proprietar şi vecin, singura făptură care ar putea să mă mai tulbure aici".

- *Back Translation*[9] *(English)*: "I have just returned from a visit to my *sulky* landlord and neighbour, the only creature who could trouble me here".

The example is selected from a novel, and the explicitation is represented by the lexical particularisation chosen. The translator prefers a semantically richer term to translate 'solitary', a term which also prepares the reader for the moroseness and the unpleasant personality of the character, rather than using the strict dictionary meaning of the source text adjective.

Although many scholars agree with these two broad categories of explicitation, more recently a distinct, richer categorisation has been proposed. According to Klaudy (2008), and favoured by other scholars such as Becher (2011*a*), explicitation is divided into four classes:

- *obligatory explicitation*, which is caused by the morphological, syntactic and semantic differences between the linguistic systems involved by the source and the target languages. It appears when some grammatical categories are present in one language and missing in the other one, and if the translator did not explicitate, an ungrammatical structure would be produced.

  At the semantic level, reality is presented in each language in a different manner, and consequently, in some cases the translator may need to bring in additional information. For instance, to translate *brother and sister* from Hungarian into English, the translator needs to explicitate by choosing either *öcs* for *younger brother*, or *húg* for

[9]This translation was made available for the present thesis only.

younger sister, or *bátya* for *older brother*, or *nővér* for *older sister* (Balázs, 2011, p. 307).

- *optional explicitation*, which is caused by differences between the source and target languages in stylistic preference. For instance, verbs with a general meaning (e.g., *'see'*, 'run') can be translated using more specific words of the target language.

- *pragmatic explicitation*, which appears due to differences or gaps in world knowledge between source and target language readers. It can also be seen as optional explicitation, with the difference that the reasons for providing additional information in the target texts are not linguistic in nature, but pragmatic, such as culture-specific words (e.g., foods or geographical places).

- *translation-inherent explicitation*, which is caused by the "nature of the translation process itself"(Klaudy, 2008, p. 107). As examples for this category are not provided by the author, this type remains unclear among scholars.

To point out the reasons behind the tendency to explicitate, it is believed to be rooted in the translator's goal of producing a less ambiguous text, which is easy to process (Robin, 2010)[10], and with a low risk of misinterpretation (Pym, 2005, p. 41). This may be because of translators awareness of their key role as mediators of messages for the target audience. Their tendency is thus to write as clearly as possible for their readers. However, they also depend on the skopos[11] of their translation.

[10]Quoted from Balázs (2011).

[11]An approach in translation studies which can be summarised by a well-known quotation: "the end justifies the means"(Reiss and Vermeer, 1984, p. 101, translated).

Returning to more recent shifts in scholars' interests, media and the explosive development of information technology have had a major influence on translation studies. Corpus-based translation studies have started challenging long-held beliefs, as they offer the means to reassess several hypotheses on the nature of translational language.

Section 2.3 below highlights recent advances in the investigation of translationese and the controversial findings regarding the validity of the translational hypotheses.

2.3 Recent Advances on Translationese, and on Simplification and Explicitation Hypotheses

This section reports the most salient studies in the literature, drawing attention to the features investigated in the research studies assessing translationese, simplification and explicitation hypotheses. It is essential to any study of translationese, or of any other translational hypothesis, to clarify which characteristics are deemed relevant for the respective phenomenon. However, the set of features investigated for a given hypothesis is a subject on which there is no agreement yet among scholars. The controversy arises because the hypotheses proposed fail to provide

This principle unveils two divergent objectives upon which the translator has to decide: the natural flow of the language, or the accuracy of the message conveyed; to concentrate on a natural flow in the target texts, or on fidelity to the source texts (Chesterman, 2004*a*).

a clear list of such features, and thus inconsistencies often occur in the literature.

Irrespective of their universality claim and controversy, a number of investigations reveal and debate interesting aspects of translational language, even though clear-cut results are still lacking within the field. Most research studies adopt the use of corpus linguistics techniques as the main methodology for the investigation of translational hypotheses.

Two directions of research are emphasised in this section: first, the corpus-based methodology largely used for the investigation of certain features suspected to stand for different universals. Second, machine learning techniques have been recently considered in the investigation of translationese. Malmkjaer (2008) argues that corpus-based techniques fit better in the search of norms of translational language rather than in the investigation of the proposed universals of translation. The latter category is scarcely used in the field, being a very recent approach to the analysis of translational language[12].

2.3.1 Corpus-based Studies

Although a degree of subjectivity in corpora compilation still prevails, investigations using corpus-based techniques are vital for Descriptive Translation Studies. This appears to be a reliable approach for ascertaining the validity of several translational hypotheses. Depending on the purpose of the study, both parallel and comparable corpora are used[13]: e.g., parallel

[12]The fundamental concepts of the discipline of machine learning are outlined in Chapter 3.

[13]More details on the notion of parallel and comparable corpora are provided in Chapter 3.

corpora are used to investigate the process of translation, to see how a message is transmitted in the target language, whilst the comparable corpora approach is suggested in product-orientated investigations of translation.

Baker (1995) argues that it is essential to use large electronic resources to investigate hypothesised manifestations in translated language, and she strongly recommends the use of corpus linguistics as a methodology. Corpus linguistics allows automatic techniques to analyse large collections of texts, carefully designed and compiled for a specific research goal. The use of monolingual comparable corpora is suggested, and she describes it as a collection of texts in the same language divided in two categories: one comprises texts translated into that language, and the other comprises "original texts in the language in question"(Baker, 1995, p. 234). She argues that the availability of these techniques represents an opportunity for the advancement of research within the domain, and further investigation regarding the distinctive features of translated text *per se* can be conducted (Baker, 1996, p. 176).

This call has had a great impact on the research community and the corpus-based approach has become one of the most important methodologies of contemporary translation studies; it was even suggested that it was the "major methodological advance associated with corpus studies"(Pym, 2008, pp. 321-322). It does not require knowledge of or familiarity with the target language and culture, and, given a set of variables, the approach has the advantage of revealing and verifying differences which intuition on its own cannot perceive (McEnery et al., 2006, p. 6). Findings based only on introspection are difficult to analyse

across various studies or languages, and thus the validation of the potential hypotheses requires more appropriate techniques of investigation, such as corpus techniques.

However, a number of arguments against the corpus-based approach were also pointed out. For instance, Tymoczko (1998) rejects the corpus-based approach as a suitable mode of research because of the *subjective judgement* of the researchers at every stage, starting from the corpus compiling process and following to the choice of the research questions and the result interpretation stage (Tymoczko, 1998). Nevertheless, research investigations on translationese and translational hypotheses are extensively based on the use of corpora: both parallel (multilingual) and comparable (monolingual) corpora. Many studies focus on the famous translation universals and attempt to bring evidence to support or reject a universal using these resources, which were carefully compiled to suit the purpose of the study.

According to Xiao and Yue (2009), existing investigations on the nature of translational language have distinct foci:

- Function-orientated descriptive translation studies focus on the impact which translation may have on the socio-cultural context;
- Process-orientated descriptive translation studies emphasise the thought process which takes place when a translator produces the new text in the target language;
- Product-orientated descriptive translation studies focus on translation as a product by comparing translated and non-translated texts in the target language.

The product-oriented branch is investigated in the present thesis, whose aim is to automatically retrieve the most informative features for the task of distinguishing between translated and non-translated texts. For this reason, in the next section, the most relevant product-oriented studies in the literature are described, pointing out the main features they used in the research investigation.

2.3.1.1 Translationese

Various studies investigate translationese as a specific feature of translational language, and several debates have arisen in the field. One point on which scholars agree is that this effect cannot be avoided in translational language (Baker, 1993; Gellerstam, 1996; McEnery and Xiao, 2007*b*). Numerous studies follow up on this phenomenon and its related translational hypotheses, most of them focusing on different universals.

The translationese hypothesis handles all the potential features that may prove to be specific to translational language in contrast to non-translational language. The formulation of translation universals, tendencies, norms or laws appeared as a result of researchers' interest in narrowing down and describing these features. Therefore, studies either discuss the manifestations of translationese in general, or they attempt to assign the features under investigation to one translational hypothesis or another[14].

[14]From this point of view, the other translational hypotheses can be referred to as translational sub-hypotheses. Throughout this thesis this term is preferred and the justification is presented in Chapter 4.

The effect of translationese is first studied in terms of *vocabulary differences*: the study analyses differences between texts translated from English into Swedish and texts originally written in Swedish (Gellerstam, 1986). The well-known publication suggests that translational language presents a statistical phenomenon which leaves its own 'fingerprints' in translation. Using a monolingual comparable corpus, comprising Swedish novels, Gellerstam (1996, pp. 56-58) reports two interesting findings: first, an unexpectedly high frequency of adjective+noun occurrences in translated texts in circumstances when an adjective on its own may be generally preferred, and second, distinct usages of reporting clauses between translated and non-translated texts. The differences of usage are assumed to be caused by the influence of the source language, English.

Lexical patterning is extensively investigated in translated texts, assuming that this feature captures a specific trait of translational language.*Repetitions* are believed to prevail in translated texts and, thus, they have been investigated by several scholars. To this end, in a study on a comparable corpus of Italian texts including native texts and translated ones, Baroni and Bernardini (2003, p. 379) conclude that translational language is repetitive, perhaps more repetitive than the language used in the native texts. Moreover, it is noted that the translated and non-translated language differ in what each tends to repeat. Analysing bigrams, they show that translations repeat structural patterns and strongly topic-dependent sequences, whilst the non-translated texts tend to repeat the topic-independent sequences (i.e., they refer to the "more usual lexicalized collocations in the language").

Tirkkonen-Condit (2002) uses Finnish texts to investigate whether human subjects can distinguish translated texts from non-translated ones and what features they use that make them think that a given text is either a translation, or an original (non-translated) text. Her results show that most of the times the subjects were unable to distinguish the translated texts, and their criteria were the frequency/scarcity of target language *specific (unique) items*. In this way, she emphasises the role of unique items in translations, bringing evidence that translations can be identified by humans only in the cases in which they can spot deviance in texts, whereas normality indicated, correctly or not, that the given text was a non-translation. Humans cannot identify texts based on their linguistic information. The only exception to this, pointed out by the author, was the frequency of personal or impersonal references in Finnish, this frequency being relevant in their decision because their usage is less frequent in Finnish compared to Indo-European languages (Tirkkonen-Condit, 2002, p. 213).

Another investigated feature is the *punctuation* used in translational language. The frequency of usage of punctuation marks is the focal point of research in an empirical study on translationese, analysing translated and non-translated texts in Spanish (Rodríguez-Castro, 2011). The results show an influence from the source texts into the target texts, a tendency to adopt the following punctuation marks: periods, commas, colons, semi-colons and em-dashes. It is reported that this tendency creates a residual effect in Spanish, indicating a lack of adherence to the appropriate stylistic conventions.

The *placement of prepositional phrases* is also investigated in translated texts, De Sutter and Van de Velde (2010) reporting experiments on texts written in Dutch and German. Using statistical techniques, the results indicate that prepositional phrase extraposition[15] occurs to a significantly lower degree in translated texts, for both languages, and syntactic differences are found. These results support the normalisation and language interference hypotheses (De Sutter and Van de Velde, 2008, 2010).

Studies on other non-Indo-European languages are also reported in the pursuit of universal characteristics of translational language, e.g. Chinese or Japanese (Xiao, 2011; Meldrum, 2009; Wang and Qin, 2010). Using a bidirectional English-Chinese parallel corpus, with literary and non-literary texts, Wang and Qin (2010, p. 164) suggest that translational Chinese exhibits: (i) more *function words* and fewer *content words*, (ii) a higher *type/token ratio*, (iii) longer *sentence segments*, (iv) "significantly more frequent use of *specific lexical bundles*".

Third person pronouns and *longer paragraphs* are interpreted to be translationese specific features in a study on Japanese investigating translated texts from popular literature (Meldrum, 2009). In contrast to an earlier observation on the ease of identification of translated texts by human subjects (Tirkkonen-Condit, 2002), the experiments in Meldrum

[15]In Dutch and German, the prepositional phrases can be placed both to the left and to the right of a verb in the subordinate clauses. Extraposition is the most common term for a structural analysis of constructions in which the prepositional phrase is shifted to the right of the verbal complex (Helmantel, 2002, p. 25). An example of this phenomenon follows: (a) PP-V: "*dat Jan [$_{PP}$ in de tuin] speelde*" (meaning "*that Jan in the garden played*") and (b) V-PP: "*dat Jan speelde [$_{PP}$ in de tuin]*" (translated as "*that Jan played in the garden*"). The prepositional phrase to the right of the verb *speelde* is said to be extraposed (Helmantel, 2002, p. 119).

(2009) suggest that readers appear to be able to detect translated texts and that their attitude towards the translationese effect appear to be neutral or slightly positive.

A set of potential translation-specific features are investigated in a Finnish corpus of children's literature (Puurtinen, 2003*b*). The resource used is a comparable corpus, comprising approximately ten million words. The source language is English and the corpus is non-lemmatised and non-parsed. The paper reports results regarding the following features: a few types of complex *non-finite constructions* (i.e., purpose, temporal and participial constructions which could be replaced by a subordinate or coordinate clause[16]), *clause connectives*[17], and *keywords*[18].

The results indicate a set of specific features of translationese: high frequency of *contracted clauses* in translations, *conjunctions* such as '*kun*'(when), '*jotta*'(in order to) appear to have different contexts and functions in translated and non-translated texts, a lack of *colloquial words* and *word forms*, and a diverse range of *reporting verbs* in translated language.

Although the main purpose of her research is the investigation of highlighted features in translations and non-translated texts, the results can also be interpreted from the perspective of translational hypotheses. Seeing simplification from the point of view of readability, Puurtinen (2003*b*) considers that the findings on the usage of non-finite constructions contradicts the hypothesis. The author emphasises that, for Finnish,

[16]An example quoted from the publication: "Mandy held her breath *expecting* her father to snap something in reply"(Puurtinen, 2003*b*).

[17]There are four categories: relative pronouns, subordinative conjunctions, coordinative conjunctions, and adverbs.

[18]Words whose frequency is unusually high or low given a certain norm, or threshold.

the use of non-finite constructions would complicate the text rather than simplifying it because of their cognitive difficulty and structural complexity. Furthermore, the explicitness level of the texts is also affected by the use of non-finite constructions: "owing to lack of connectives", several relations between propositions remain implicit in the texts. In terms of the usage of connectives in translations, the results fail to clearly support or contradict explicitation (Puurtinen, 2003*b*).

In terms of analysing the *part of speech classes* in translational language, a remarkable study was conducted by Borin and Prütz (2001). Translated texts from Swedish into English are under investigation, and the research focus is on the syntactic dimension of translationese, unlike most of the studies which investigate the lexical dimension. The overuse and underuse of part of speech n-grams are analysed, and an *overuse of adverbs, infinitives, pronouns and sentence-initial prepositions* is pointed out.

As most of the studies interpret their findings from the well-known translation universals' perspective rather than the more general hypothesis, translationese, the following subsections describe the most important results reported in the literature regarding simplification and explicitation.

2.3.1.2 Simplification

Scholars within the domain have attempted to investigate various possible features to validate the simplification universal. Different aspects of the hypothesis have been both criticised and sustained. The hypothesis remains rather controversial since there is a lack of consistent findings. Moreover,

it appears to share some of the attributes in favour of explicitation. For instance, *sentence length* is one attribute investigated for both universals, just as *lower lexical density* is seen as evidence that texts are easier to understand, simpler, and thus more explicit.

Simplification is a hypothesis which can be seen as the translator's tendency to improve *readability* (Puurtinen, 2003*b*; Corpas, 2008), which leads translators to employ various methods of improving the comprehensibility of their texts for the target audience. The translators' preferences for *disambiguation* and simplification of the message conveyed was remarked earlier in the literature by Vanderauwera (1985) in a study in which the results show that *ambiguous pronouns* are mapped onto precise forms and, thus, the correct understanding of the message is conveyed in the target text: "where quotation marks fail to distinguish a person's speech or thought in the source text, they are almost invariably restored in the target text"(Vanderauwera, 1985, p. 94).

Evidence for lexical and syntactic simplification is brought by various studies, such as Laviosa (1998) or Olohan and Baker (2000) for English translations, or Xiao et al. (2010) for Chinese translations. One of the best-known corpus-based studies on simplification is by Laviosa (1998, 2002), who conducted a set of experiments on translations into English[19]. Her findings relate to simplification at lexical level and reveal the following outcomes (Laviosa, 2002, pp. 60-62):

[19]The Translational English Corpus was a 10-million-word resource at that time, reaching a size of 20 million words by the year 2001 (Xiao, 2010*b*, p. 3). Details of this resource can be found in Chapter 3.

- translations have a relatively *lower* percentage of *content words*[20] *vs. grammatical words* (i.e., lower lexical density);
- translations exhibit a relatively *high proportion of high-frequency words*;
- translations have a relatively *great repetition of the most frequent words*;
- translations have a *lower ratio of lexical to running words*[21];
- translations have a *lower average sentence length*;
- translations exhibit *less variety in most frequently used words* (fewer lemmas).

She concludes that corpus-based research allowed translation universals to be better defined, to progress from manual investigations to large scale, target-oriented research, to provide more consistent evidence, and to consider a wider range of socio-cultural factors (Laviosa, 2002, p. 75). Despite these findings, the author concludes that a hypothesis about potential universal features specific to translational language cannot be determined based only on the experiments conducted (Laviosa, 2002, p. 51).

In a similar manner, using the ZJU Corpus of Translational Chinese, a comparable corpus of translated and non-translated texts, studies have been carried out which indicate that the core patterns of lexical features

[20]By 'content words' is understood the lexical words found in texts.
[21]Metric referred to as the information load (Olohan, 2004, p. 100).

presented by Laviosa (1998) are also supported in different languages. For instance, Xiao et al. (2010) provides evidence that:

- translational Chinese exhibits lower *lexical density*;
- in translated texts there is a low proportion of *lexical words* over function words;
- a higher proportion of *high-frequency words* over low-frequency words is found in translational language;
- a *high repetition rate for high-frequency words* is noted as characteristic of translational language.

The *connectors* used in translational language appear to be simpler and they are more frequently found in translations than in non-translated texts. These results bring evidence for both simplification and explicitation. In addition, their experiments shed light on normalisation. The results concerning the *use of passives* indicate that the source-induced difference between translations and non-translations suggests that normalisation may be language-specific and it does not hold for Chinese.

Punctuation is another feature investigated for simplification. Besides being used in the investigation of explicitation, punctuation is also suggested as a feature of simplification at a stylistic level. Malmkjaer (1997) finds that in English translations punctuation appears to be stronger than in non-translations: *commas* are replaced with semi-colons or full stops, whilst *semi-colons* are replaced with full stops. The author thus interprets that the longer sentences from the source texts are broken up

into shorter and simpler ones in the target texts, which results in a higher readability level in translated texts. Baker (1996) views these traits as a strategy of simplifying and disambiguating the message in translations. However, interpretation can also head towards normalisation. Studies such as Vanderauwera (1985) and Malmkjaer (1997) provide evidence for shifting norms towards a more normalised punctuation in translational language.

In contrast with the observation on *shorter and simpler sentences* in English translations (Malmkjaer, 1997) is the finding on average sentence length presented by Laviosa (1998). Analysing English texts, Laviosa (1998) finds that in translated texts the mean sentence length is higher than in non-translations. Also, in Chinese translations, the results on sentence length show a higher mean value for translations (Xiao and Yue, 2009).

Consequently, the sentence length feature appears to be showing controversial findings. In addition, Xiao et al. (2010) suggest that this feature may be a language- or genre-dependent attribute: for genres such as humour, the mean value is lower than in academic prose, where longer sentences are predominant.

Another feature analysed in translated texts is the *vocabulary range*. Blum-Kulka suggests that *lexical simplification* appears in "the process and/or result of making do with less words"(Blum-Kulka and Levenston, 1983, p. 119). The expectation is that lexical density is rather low among translations, and Baker suggests that it should be interpreted as a "subconscious strategy of simplification"(Baker, 1996, p. 183). However, the subconscious claim, either regarding simplification or other hypothesis, has given rise to debates among scholars (Becher, 2011*a*).

Several corpus-based studies investigate the lexical level of translational language in comparison to non-translational language. Baker (2004) analyses *recurring lexical patterns* in English texts, such as *"in other words"*, *"at the same time"*, and temporal and spatial phrases, like *"in the middle of"*, *"for the first time"*, etc. The findings show that translations exhibit a higher frequency of these types of lexical phrases in comparison to non-translated texts. Also, the distribution of the lexical phrases across the texts appears to be less even in translational language.

A few experiments also employ natural language processing tools, such as part of speech taggers or parsers, in conjunction with statistical metrics, such as the t-test (Corpas, 2008; Xiao et al., 2010). The adoption of natural language processing tools is emphasised and the importance of using statistical metrics in the investigations of translational hypotheses is highlighted (Corpas, 2008, p. 172). Consequently, a set of experiments tested the statistical significance of a range of features in a large medical and technical comparable corpus in Spanish (Corpas, 2008). The findings show that simplification seems to be validated for *lexical richness*, and contradicted in terms of *complex sentences, sentence length, depth of syntactical trees, information load, and ambiguity*[22].

Additionally, another study focusing on simplification from the point of view of readability (Corpas, Mitkov, Afzal and Pekar, 2008) found that translated Spanish texts appear to exhibit lower *lexical density and richness*, seem to be more *readable*, have a smaller proportion of *simple sentences* and appear to be significantly *shorter*, whilst *discourse markers* are used significantly less often. Simplification has been spotted in technical

[22]Where ambiguity is computed as the average number of senses per word.

translations in comparison to medical ones and seems to show that texts written by semi-professionals do not have such simplification traits.

The study also investigates the convergence hypothesis from the similarity point of view, i.e., to what extent a text is similar to another one (Corpas, Mitkov, Afzal and Pekar, 2008; Corpas, Mitkov, Afzal and García, 2008; Corpas, 2008). The focus of the study is on the following research objective: to what extent translated or non-translated texts converge on the basis of selected parameters, namely type/token ratio, lexical density, sentence length, use of simple and complex sentences, use of aspect, discourse markers, and conjunctions. In the research conducted by Corpas, Mitkov, Afzal and Pekar (2008), as well as in the experiments reported by Corpas (2008), it is demonstrated that both the convergence and simplification hypotheses are contradicted.

In a monolingual Finnish subcorpus, extracted from the Corpus of Translated Finnish, simplification traits are found at various levels: translational language showed a *simplified discourse*, lower *lexical density*, a high proportion of *high frequency words*, a lower frequency of *hapax legomena*[23] (Nevalainen, 2005).

Although some corpus-based studies bring evidence for the existence of such a phenomenon (Laviosa, 2002; Nevalainen, 2005; Xiao et al., 2010), it is still remarkably challenging to reach agreement regarding the validity of this hypothesis and regarding the features which characterise the simplification hypothesis.

[23]Words that occur only once in the corpus.

This hypothesis is known to be a controversial claim, as there are various studies bringing evidence both for and against it. It is contested by studies on collocations (Mauranen, 2000), lexical use (Jantunen, 2001), and syntax (Eskola, 2002; Jantunen, 2004*a*; Corpas, 2008).

However, it is important to note that the studies investigating this hypothesis cannot always be compared. Laviosa-Braithwaite (1996) emphasises this point, stating that the early findings on simplification can be incoherent and the results reported for different datasets, having various research questions, cannot be compared (Laviosa-Braithwaite, 1996, p. 534). Also, the methodology used has an important role in the interpretation and comparison of findings: for instance, Laviosa (2002) compares features in terms of frequencies in translations and non-translations, whilst Jantunen (2001, 2004*b*) analyses a selection of individual items (i.e., the synonymous degree modifiers *hyvin, kovin and oikein*, all of them roughly meaning *'very'*). Hence, their findings are not comparable in terms of methodology, a statement also supported by Mauranen (2008, p. 40).

Perhaps the major drawback in the investigation of simplification and the main cause of disagreement in the interpretation of the results is due to ambiguity in defining and quantifying the very concept of 'simple' in texts. Although some features are suggested and interpreted as paving the way to a simpler, easier-to-read text, there are cases when a feature supposed to ensure simplicity can also cause a certain difficulty at the text level. For instance, simplifying the structure of the sentences into more simple main clauses with a few subordinates may also result in a certain complexity at

text level: a reduced coherent flow and an impression of fragmented and hard-to-follow texts (Mauranen, 2008, p. 40).

Similar ambiguities in terms of defining and quantifying translational hypotheses are issues which need to be addressed and analysed within the domain. The most important research studies on explicitation follow in the next section.

2.3.1.3 Explicitation

The other proposed translational hypothesis relevant to this thesis, explicitation, is probably the most studied hypothesis in translational hypotheses research. In contrast to simplification, the explicitation hypothesis is perhaps the least controversial among scholars.

The hypothesis states that "explicitation is a universal strategy inherent in the process of language mediation "(Blum-Kulka, 1986, p. 21). Baker includes it among the universals, assuming that professional translators have a tendency to "spell things out rather than leave them implicit"(Baker, 1996, p. 180).

Considering the attempts at defining this phenomenon, there are three claims regarding this hypothesis. First, a potential explicit characteristic is assumed to appear in translational language. Second, it is hypothesised that this characteristic is independent of language-specific explicitness (Blum-Kulka, 1986, p. 19). Third, it is postulated that this characteristic of translational language appears as a consequence of the process of translation (Blum-Kulka, 1986, p. 21).

All these claims are analysed within the domain, and for the first one several findings are reported in favour of this hypothesis. However, independence from the language-specific explicitness is still a challenge; moreover, it depends on the conclusion drawn from the first claim. Given the current definition of the hypothesis, it is sufficient to find one language which brings consistent evidence against it. For the third claim, explicitation seen as a translation-inherent strategy, there are scholars who strongly argue against it, seeing it as unmotivated (Becher, 2011*a*).

The explicitation hypothesis is confirmed by several researchers, and an overview of the features typical of explicitation, largely suggested by Gumul (2006), is illustrated as follows:

- adding connectives: e.g., 'and', 'so', 'thus', 'although', 'despite' (Blum-Kulka, 1986; Puurtinen, 2003*a*, 2004);

- shifts from vaguely to more explicitly cohesive: e.g., '*and* they decided to wait' becomes '*so* they decided to wait' (Øverås, 1998);

- adding modifiers and qualifiers: e.g., 'a *strange* mixture', '*serious* psychological damage' (Vanderauwera, 1985);

- disambiguating, filling out elliptical constructions: e.g., 'some of the other consequences, and there were many of them, some *[of the consequences were]* very important' (Øverås, 1998; Pápai, 2004);

- adding a proper name to a generic name: e.g., 'every *American* citizen' (Øverås, 1998);

- replacing general words with more specific ones (lexical specification): e.g., 'say' becomes '*accuse*' (Øverås, 1998; Perego, 2003; Klaudy and Karoli, 2005);

- disambiguating metaphors: e.g., 'they stand in a frightened lump' becomes '*they stand huddled together*' (Øverås, 1998);

- including additional explanatory remarks: e.g., '*the web page* liberty.org' (Baker, 1992; Pápai, 2004);

- repetitions, shifts from reiteration as paraphrase in ST to reiteration as partial or identical repetition in TT: e.g., 'a bridge across the Thames (...) the north bank of *the river*' becomes 'a bridge across the Thames (...) the north bank of *the Thames*' (Øverås, 1998; Gumul, 2004);

- syntactic expansion: 'the girl I saw' becomes 'the girl *that* I saw' (Olohan and Baker, 2000);

- disambiguation of pronouns, shifts from referential cohesion to lexical cohesion: e.g., 'overlap between *them*' becomes 'overlap between *these sections*' (Olohan and Baker, 2000; Pápai, 2004);

- adding lexical information that belongs to common knowledge for the source text readers but is not directly available to target text readers: e.g., '*Huesca*' becomes *'the city of Huesca in northern Spain*' (Pym, 2011);

- reformulation markers known as discourse markers indicating equivalence: e.g., 'that is to say', 'namely', 'in other words', 'that is to say', 'to put it differently' (Baker, 2004, 2007; Xiao, 2011);

- repetition of names and noun phrases in preference to the use of pronouns (Olohan and Baker, 2000).

Considering the above enumeration, Becher (2009) summarises the key notions surrounding the explicitation phenomenon as follows:

- *Coherence*: A discourse is coherent if and only if all of its elements directly or indirectly contribute to the discourse purpose (Grosz and Sidner, 1986). In the study of translation, the term is introduced by Reiss and Vermeer (1984) and they classify coherence into two categories: first, the *intratextual coherence*, which means that "the message introduced by the translator must be interpretable in a way that is coherent with the target recipient's situation" (Reiss and Vermeer, 1984, p. 113, translated). The second category is the *intertextual coherence*, referring to the coherence which exists between the target text and the source text (Shuttleworth and Cowie, 1997)[24].

- *Cohesion*: A discourse is cohesive if and only if it contains formal cues/markers, such as connectives, that signal its coherence (Bublitz, 1998). Cohesion is a form of explicitness (House, 2004).

- *Connective*: A conjunction, sentence adverbial or particle that creates cohesion by assigning thematic roles to sentences, such as cause-effect (Pasch et al., 2003).

[24]Intertextual coherence, also known as *fidelity*, is considered to be present if consistency across the following three rules is achieved: (i) the original source text message, (ii) the way the translator understands and interprets this message, (iii) the manner in which the translator conveys the message to the target text reader (Shuttleworth and Cowie, 1997).

Among the first studies investigating this hypothesis, traits of explicitation are supported in terms of *supplementary explanatory phrases*, *expansion of condensed passages*, *resolution of ambiguities*, *use of repetitions*, and *cohesive devices* in translational language (Vanderauwera, 1985; Blum-Kulka, 1986).

Considering the classification of universals illustrated by Chesterman (2004*a*) and presented earlier in Section 2.2.3.1, the explicitation hypothesis naturally falls into the category of S-universals, since shifts between source and target texts and additional information occurring in the target texts are analysed. However, it can also be investigated as a T-universal, when the explicitness level of translational language is compared to the degree observed in non-translational language. The former situation requires a parallel corpus, whilst the latter relies on a comparable corpus. There are a few studies which investigate this hypothesis from both points of view: as an S-universal, and as a T-universal. For instance, Chen (2006) researches explicitation at both process-level (and thus, as an S-universal) and product-level (as a T-universal), investigating to what extent connectives are explicitated in Chinese.

An analysis of connectives, specifically *conjunctions* and *sentential adverbials*, is reported and experiments are conducted on a corpus which comprises English source texts and their two independent translations into Chinese. As a reference corpus, a comparable component is used comprising native Chinese texts in the genre of popular science. The results support the explicitation hypothesis both at the process-level, when comparing to source texts, and also at the product-level, when comparing translations to non-translations.

On English-Norwegian and Norwegian-English literary translations, Øverås (1998) manually annotated all the explicitations and implicitations found in 1000 sentences extracted for each translation direction. The study reports the frequency rates regarding the number of explicitation shifts focusing on *lexical cohesion*[25] and grammatical ties, such as conjunctions and references. She confirms the hypothesis by observing an increased cohesion in translational language.

As already pointed out, *sentence length* is also investigated in the pursuit of voluntary explicitation, not only for simplification hypothesis. Using a small-scale bi-directional corpus of Portuguese and English source texts and translations, Frankenberg-Garcia (2004) suggests that the reason for the increased number of words found in translated texts is highly likely to be because of the differences that appear between the source and the target texts.

More recently, the explicitation hypothesis has also been analysed in terms of *reformulation markers* (Xiao, 2011). In a study investigating translational Chinese language, it is suggested that "reformulation markers function as a strategy for explicitation in translations, which tend to use oral, stylistically simpler forms than non-translated texts" (Xiao, 2011, p. 145).

However, this hypothesis also has its controversial aspects. Besides arguing against the subconscious nature claimed by Blum-Kulka's, Becher (2011*a*) questions the findings yielded in the investigations on explicitation

[25]The concept of cohesion is based on the definition provided by Halliday and Hasan (1976), according to which cohesion is categorised into two classes: grammatical cohesion (reference, substitution, and ellipsis), and lexical cohesion. Conjunction is a type of cohesion considered to be on the borderline between the two, mainly grammatical, but with a lexical element to it (Halliday and Hasan, 1976, p. 6).

in general, mainly because of the methodology adopted and the interpretation of the results. In particular, he analyses closely the studies conducted by Olohan and Baker (2000) and Øverås (1998). The author states that these studies "fail to provide conclusive evidence" because of a set of methodological drawbacks, such as: potential explanations other than the suggested explicitation hypothesis are not sought (i.e., such as other types of explicitation, source language interference, the effect of other potential translational hypotheses) and several studies do not consistently adhere to the definition of explicitation (Becher, 2011*a*, p. 57). The author emphasises that explicitation is a highly complex phenomenon in translations and it requires a more rigorous investigation than it has received to date.

Besides these controversial aspects, and even though explicitation is known to be the least controversial among the universals defined, there are also studies which bring evidence against its presence among the universals of translation, and thus the implication that the hypothesis is true irrespective of the language pairs involved is disproved.

At this point, an important observation is noted. In the present thesis, the *"universal"* terminology implies, by definition, that the assumed phenomena occur in every translated text, in every act of translation, regardless of the source or target language. The appropriateness of the "universal" terminology is thus doubted, and for this reason, this thesis sustains the adoption of the "translational hypotheses" terminology rather than "translation universals", and respectively, of the "translation hypotheses research" terminology rather than "translation universals

research". The preference for the "hypothesis" terminology is justified earlier, in Section 2.2.3.

Returning to studies against explicitation, the empirical study on a parallel corpus of English-Korean translation described by Cheong (2006) investigates target text expansion as a trait of explicitation and target text contraction as a trait of implicitation. To this end, four parameters are analysed for this complex phenomenon: the word count rate for the overall text length, the connectives frequency rate to assess the changes at the cohesive level, the parenthesis frequency rate for the supplementary explanatory information, the bracket frequency rate to locate repetitive orthographic representations of the same information.

Although the word count rate proved to be useful in tracking the changes in the source and the target texts, this feature alone has its drawbacks, since an expansion in one text segment can be counterbalanced by a contraction in another one. Furthermore, the use of connectives, acknowledged in the literature to be in favour of explicitation, appears to be against these findings by representing a contraction in the English-to-Korean dataset, and thus supporting implicitation. The author suggests that the reason for this drop can be one of the linguistic features of Korean itself (i.e., the cohesion of the texts does not only rely solely on connective adverbs, but also on auxiliary suffixes attached mostly to nouns, which indicate case marking or clause connections). The third feature analysed, the parenthesis frequency rate, is explained as the need for translators to enhance the accessibility of the target text, whilst the use of brackets can be a result of the differences in text production conventions between the two languages involved.

The study concludes that the explicitation phenomenon is not necessarily present in all of the translations, and in addition, an observation is noted: "the direction of translating language combination leads to a different target text expansion/contraction behaviour of texts"(Cheong, 2006).

In addition to the corpus-based research dedicated to translationese and translational hypotheses, a few studies adhere to a distinct perspective, granted by machine learning algorithms. These studies are described in the following subsection.

2.3.2 Machine Learning Approach

Machine learning has been rarely adopted within translation research[26], as only a few scholars have undertaken this approach in their analysis of translational language. Unlike the classical directions of research in the study of translationese, by means of pilot studies or by adopting the preferred methodology since the nineties – the corpus-based approach – a different perspective provided by the use of machine learning techniques has been recently noted (Baroni and Bernardini, 2006). A machine learning approach is reported for the task of classifying Italian texts as translated from other languages or originally written in Italian. The study is the closest research approach to the experiments illustrated in the present thesis[27].

[26]As the present thesis adopts this approach, a brief introduction to the discipline of machine learning is outlined in Chapter 3.

[27]The main differences are highlighted in Chapter 5, Section 5.4.1.

The resource used in this study is a monolingual comparable corpus comprising geopolitical journal articles written in Italian. The comparability is assessed between translated and non-translated text types. The research study relies on an SVM classifier and on the use of n-gram features. The feature vector[28] represents a text in terms of unigrams, bigrams, trigrams, and word forms, lemmas, part of speech tags, and mixed, respectively (Baroni and Bernardini, 2006). In other words, the characteristics investigated in the study are: combination of unigrams, bigrams, trigrams, and word forms; lemmas; part of speech tags; and other parameters which consider various mixtures of the previous characteristics mentioned. It is reported that the SVM classifier depends heavily on *lexical cues*, the *distribution of n-grams of function words and the morpho-syntactic categories* in general, and on *personal pronouns* and *adverbs* in particular. Their findings show that relatively shallow data representations[29] comprising the indicators mentioned can be sufficient to automatically distinguish professional translations from original texts with an accuracy of up to 86.7%.

The SVM algorithm was also employed by van Halteren (2008), who applied it to the task of identifying the source language of a given text from the Europarl corpus[30]. The Europarl corpus is a parallel corpus comprising texts written in twenty-one European languages (Koehn, 2005), six of which – English, German, French, Spanish, Italian and Dutch – were used in these experiments. Their study aims at identifying the source language of the texts from the point of view of natural language processing, rather than

[28]This concept is further explained in Chapter 3, Section 3.3.

[29]The concept of data representation as well as other important machine learning concepts are defined in Section 3.3.1.2.

[30]`http://www.statmt.org/europarl/`

from the perspective of translation studies. However, it also represents a research contribution for translational hypotheses research, not only for machine translation.

The other two methods used by van Halteren (2008), besides the SVM technique, are as follows: one implies the use of language markers classification (i.e., the n-gram which occurs with a higher frequency for a certain source language L than all the other source languages is considered a marker for L) and the linguistic profiling classification (i.e., the classification relies on the underuse and overuse of specific n-grams). All the methods employed analyse *specific words*, *n-grams*, not part-of-speech classes. Their system obtains an accuracy of 87.2% to 96.7% in the task of identifying the source language when all of the six classes are trained. As specific words are used in their data representation, the results are influenced by vocabulary, discourse structure and probably syntax, according to their explanations. Also the behaviour patterns of parliamentarians of distinct countries, as well as the contrast between the source and target language, contribute to these results.

Influenced by Baroni and Bernardini (2006), Kurokawa et al. (2009) prove that it is possible to automatically detect the direction of translation and they explore mixed text representations by combining function words and part of speech tags instead of content words, or by employing words, lemmas, and part of speech tags in their model. The resource used is the English-French Canadian Hansard, and their aim is to classify text chunks and sentences as original versus translated. The SVM algorithm achieves an accuracy of 90% for the chunks, a value obtained based on n-gram words, and 77% for the sentences. They further analyse how these

differences between texts can impact the SMT system, and they conclude that considering the directionality when training the SMT system can influence the quality of the output.

In the same direction of research, whose aim is to identify the source language of translated texts using machine learning, a few studies are reported in the context of statistical machine translation (Koppel and Ordan, 2011; Lembersky et al., 2011; Volansky et al., 2011; Lembersky et al., 2012). It is shown that machine translation systems based on translated language models outperform systems which employ language models based on original texts. These studies corroborate and strengthen the findings reported in the published experiments related to the present thesis.

Using function word frequencies to represent chunks of text, Koppel and Ordan (2011) use the Bayesian logistic regression to identify translated and non-translated texts. Their high accuracy results based on the Europarl corpus provide evidence for translationese and interference hypotheses. Although both van Halteren (2008) and Koppel and Ordan (2011) use the Europarl corpus, their results may sustain interpretationese, as the type of texts included in the corpus, a dedicated resource for SMT[31], are transcriptions of speeches delivered during the European Parliament meetings.

Furthermore, focusing on language models and training various 4-grams language models[32] on various corpora from Europarl, Lembersky

[31] `http://www.statmt.org/europarl/`.

[32] They use the SRILM toolkit reported by Stolcke (2002). A *language model* is a type of statistical model which estimates the prior probabilities of word strings (Stolcke, 2002).

et al. (2011) bring evidence on translationese, and emphasise that theoretical translation studies hypotheses can contribute to the performance obtained by the SMT systems.

In the past few months, another statistical algorithm, also used in machine learning, has been reported in the investigation of Dutch translations: the logistic regression model. The Belgian Dutch component of a ten-million-word Dutch parallel corpus is used in a research study conducted by De Sutter et al. (2012). The parallel corpus comprises translations in Dutch and their source texts in French and English. Text types include the following: fictional and non-fictional literature, journalistic, administrative, and instructive texts. Exploring a set of different *lexical features* and eight different Dutch varieties, also called lects[33], they create a multidimensionality which requires statistical techniques. To this end, they analyse their data by combining the logistic regression model with an advanced form of correspondence analysis.

Their experiments analyse whether translators prefer formal lexemes to neutral ones in their translations. They use ten lexical alternation variables (profiles), each of them consisting of one formal variant combined with one neutral one. More precisely, the ten profiles are ten sets of synonymous naming variants used to express the same concept (e.g., 'car' versus 'automobile'). For instance, an example of formal variant is 'numeral+*maal*', whilst 'numeral+*keer*' is neutral. Both 'maal' and 'keer' are translated as 'times' in English. To determine the degree of formality

[33]The authors use the term *lect* to refer to the five types involved (fictional and non-fictional literature, external communication, journalistic texts, instructive texts and administrative texts), and the other three varieties being translations from English, translations from French, and non-translated texts.

found in texts, the profile-based approach is adopted, charting the naming preferences for each variety. The binary logistic regression is employed to predict the formality variation. The algorithm has two classes: the neutral lexeme class, and the formal lexeme class. The feature vector contains two variables: the text type and the source language.

The results show that, in terms of formal lexemes, significant differences between translations and non-translations do occur and that these differences are dependent on the source language (De Sutter et al., 2012).

In the next section, the strengths and drawbacks of the research on translationese and related hypotheses are illustrated.

2.3.3 Strengths and Shortcomings

In the past two decades, there has been increased interest in the analysis of the nature of translational language, as scholars focused on the investigation of translationese and the well-known translation universals. Their findings shed light on some of the statements reported, but they also prove to be difficult to interpret: i.e., whilst translationese seems to be accepted as unavoidable within translational language, the results on universals appear to be intriguing, difficult to interpret, and sometimes controversial.

Perhaps the most important strength of these universals lies in their explanatory power, in their potential to capture tendencies within translated texts which will raise awareness among professionals about the potential effects which appear in translational language. In this way,

the theoretical background could pave the way for further developments towards more accurate translations.

Also, a set of drawbacks can be observed in the literature to date. First, it is noted that the hypotheses, as they are currently explained and formulated, contradict each other. For instance, if the untypical lexical patterning hypothesis is proven true, this tendency will contradict the simplification hypothesis by the following rationale: according to simplification, there is an overuse of the most typical structures and words of the target language, whilst the untypical lexical patterning tendency precisely suggests that occurrences of untypical signs also exist (Chesterman, 2011). Another example of such a contradiction is the lengthening tendency assumed to be specific to translations, which appears to run counter to the simplicity hypothesis, since long sentences are known to pose a higher degree of complexity. Yet, both of the assertions might be true at the same though if, for instance, there are to be found long and easy to understand sentences occurring with a high frequency in translated texts.

According to Pym (2008), Toury's laws can also be seen as a contradiction because of the following: if all the translations are alike (according to the standardisation law), then it is unlikely for them to be like their source texts (according to the interference law), basically, like their "very individual and distinct source texts" (Pym, 2008, p. 314). On this occasion, Pym (2008) suggests that a potential unification between Toury's laws and Baker's translation universals is probably required.

As a consequence of the contradictions above, clear-cut borderlines between the hypotheses formed are not well defined. These claims seem to lack logical coherence, which should definitely exist since all of them investigate and postulate tendencies regarding the same translational language. Considering that the hypotheses themselves contradict each other, it is expected to have controversies and difficulties at the analysis stage or when interpreting the results. The theoretical background which allows several points of view, being maybe too general, is likely to raise questions at each research stage: from the hypothesis itself until the evaluation and the interpretation of the results. Various debates regarding the validity of the hypotheses continue, and their logical coherence is questioned across the scholarly community (Bernardini and Zanettin, 2004; Becher, 2011*b*).

More objections regarding translationese and related translational hypotheses are reported within the field:

- the need for a unified theoretical basis for the domain is being pointed out, and a call for the potential unification between Toury's laws and the well-known translation universals is made by Pym (2008).
- the translational hypotheses formalised within the domain lack precise and clear criteria through which a hypothesis might be confirmed.
- the investigation of the opposite effect or features: e.g., if one study finds that feature X appears with a higher frequency in texts of type A, then it should be expected that on the same dataset the opposite feature, namely -X, has a lower frequency on the same type of texts

A. Nevertheless, few studies employ the use of opposite indicators in their investigations.

- terminological chaos: although translation studies is a field that has developed considerably in the last twenty-five years, there are several controversies regarding the terminological use. Although a few scholars undertake this task by creating appropriate resources (Pym, 2011), probably more restrictions in the definition of their hypotheses and concepts would generate fewer terminological debates.

- the tendency to overgeneralise trends from case studies (Gentzler, 1993): typical cases are important qualitative studies needed within the domain. However, their evidence for hypotheses which claim universality over a certain trend is definitely questionable.

- generalising trends based on studies on specific genres or languages: in other words, if a feature F, which is in favour of a hypothesis H, is found to be specific to translated texts written in a language L, then that feature F is only a candidate for a feature specific to general translation. Further studies on distinct languages and distinct language families should be conducted before concluding that the hypothesis H, or the feature F, holds true for or is characteristic of all translations.

- the findings yielded from distinct methodologies cannot be compared (Laviosa-Braithwaite, 1996; Mauranen, 2008): e.g., features in terms of frequencies across large datasets cannot be compared to the findings yielded by the analysis of individual items (such as investigating the translations of the verb 'tell').

- a highly-debated aspect is the universality issue: it has been pointed out that the investigation of all translations from all languages from all times is highly unlikely (Tymoczko, 1998).

As mentioned above, considerable attention is drawn to the 'universality' aspect implied by these translation hypotheses. Whereas scholars like Tymoczko (1998) or Bernardini and Zanettin (2004) argue for the 'universality' element, others, such as (Toury, 2004), consider that the value of these hypotheses stands in the explanatory power of the linguistic system used in translation. Even though Toury (1995) accepts the "universal" terminology, using it in an earlier publication (Toury, 1979), he adopts the 'law' nomination for the hypotheses he proposed, implying a less deterministic claim in this manner. For the same reason, the term 'hypothesis' is preferred throughout this research.

On the one hand, the assertions require adequate practical support to be validated even without considering the universality aspect. To this end, a more rigorous methodology that can bring novel perspectives to these controversial statements is necessary. On the other hand, the universality characteristic is a controversial subject on its own, as this term has a wide coverage: it involves all the translations into all the languages available and all of them should manifest the same feature or set of features according to the translationese effect or translation universals. The evidence provided only for a few specific languages is thus insufficient to accept the universality implication. It is enough to bring evidence against these statements just for one language to suggest that the claims need to be refined.

Nevertheless, the quest for universals has its strengths. As Chesterman (2004*b*) emphasises, these universals appear as a stage on their journey to "high level generalisations". Although still controversial, these hypotheses are appreciated for their possible explanatory power (Toury, 2004), and current investigations show that translational language may share specific features in comparison with non-translational language, but these characteristics prove to be complex, impure, and challenging to discover.

2.4 Proposed Line of Research in this Thesis

As the previous section outlined, there are several important corpus-based studies investigating translationese, simplification and explicitation. Most of these studies are small-scale experiments, analysed manually or semi-automatically, which consider texts from a given pair of languages: certain patterns have been found in translations from language A into language B. However, a few aspects arise in these investigations, such as: first, these studies tackle only one hypothesis at a time, although distinct hypotheses occur at the same level, translational language. Second, for each tendency analysed, the research studies investigate a short set of features proposed within the literature, and thus have a limited perspective on other features which, if considered, could lead to relevant patterns.

A third important aspect regards methodological preference, which is clearly for the corpus-based direction; only a very low number of scholars adhere to other research approaches. Machine learning algorithms are rarely used within the domain, although they may have the potential to uncover interesting findings about translational language. Moreover,

the methodological preference could be selected according to another important aspect, which is highly discussed within the literature, namely the universality factor implied by translational hypotheses. Currently, the universality factor does not seem to have a rigorous methodology to be validated within the field yet, because the analysis of all the translations from all the languages is unlikely to be feasible. Therefore, a multilingual approach is desirable within the field.

The fourth aspect is the lack of an ability to compare the findings across studies; this factor is closely related to the approach chosen. In order to allow for the findings to be comparable across different studies, an appropriate methodology should be employed.

Also, another aspect to consider is that, to the best of the author's knowledge, no study has yet been able to provide a ranking of the most influential features which are able to identify translational language from non-translational, paving the way to a better understanding of the nature of translational language.

Considering all these aspects, a multilingual methodology which handles a rather large set of features all at once, on the same dataset, and is able to rank all these features according to their influence on the overall system, would be required.

The present thesis addresses these issues and reports a learning model able to analyse and rank different features of translational language, providing a set of patterns extracted in the learning process. Also, it emphasises the importance and the potential of the machine learning approach to translational hypotheses. As machine learning systems are

scarcely reported in the literature, the main core concepts of such a framework are briefly pointed out in Chapter 4, before describing the data representation used in the current experiments.

2.5 Conclusions

The purpose of this chapter is to highlight the main directions of research within descriptive translation studies, emphasising the most relevant studies involved in the literature.

The preliminary theoretical background is introduced in Section 2.2, establishing the current study as an interdisciplinary investigation focusing on three areas: translation universals, machine learning and natural language processing. Different categories of translational hypotheses are then presented, focusing on the relevant hypotheses to the present thesis, namely translationese, simplification and explicitation. These translation universals have a great impact in the research community, being widely investigated in the last two decades.

Section 2.3 describes the main approaches used in the literature: there are several corpus-based studies and a few investigations adopting machine learning algorithms. The section ends by pointing out the overall strengths and drawbacks of the existing research, the chapter continuing with the proposed line of research of the current thesis in Section 2.4. The direction of research is contextualised within the descriptive translation studies domain and the need to adopt a machine learning perspective in the investigation of translational language is justified.

The next chapter introduces the resources needed for this research: comparable corpora for Spanish and Romanian, and the machine learning tool used for the present experiments.

Chapter 3

Resources Required

3.1 Overview

The line of research reported in this thesis connects various concepts and techniques from different fields: it applies machine learning algorithms to the investigation of the nature of translational language. As a result, it is necessary to briefly introduce the background concepts, tools and resources relevant to this research.

This chapter outlines the necessary information regarding the resources and tools used for this research. Since the hypotheses studied compare translational language to non-translational, the type of resource adopted is comparable corpora.

Section 3.2 introduces a basic type of resource in the domain of Corpus Linguistics, which is also widely used in Natural Language Processing: the concept of the corpus, and two subtypes relevant to this research: the comparable corpus and the translational corpus. This outline is

provided because the research reported in this thesis comprises two sets of experiments that both use monolingual translational comparable corpora. These corpora include texts written in two languages, Spanish and Romanian. Details on the Spanish corpus are presented in Section 3.2.2.1, whereas the description of the Romanian corpus, RoTC, is outlined in Section 3.2.2.2.

Afterwards, Section 3.3 outlines the core concepts of the discipline of machine learning and introduces Weka (Bouckaert et al., 2012), the machine learning toolkit extensively used in this research. This software requires certain pre-processing stages to be able to apply its machine learning techniques to the investigation of translational language.

3.2 Translational Comparable Corpora

The investigation of the translational hypotheses proposed in the last two decades relied on the availability of certain resources. Most of these hypotheses (e.g., translation universals, laws or norms) imply a comparison between translated texts produced by professional translators and non-translated texts. As a consequence, there is a need for monolingual comparable corpora specifically designed for the study of translational language. Briefly described, these corpora need to contain two subcorpora: a subcorpus that comprises translated texts, and a comparable one which comprises non-translated, original texts.

The organisation of this section is as follows: first, several reasons are given as to why it is important to compile comparable corpora for

translation studies, and second, the notions required for this study are illustrated. Section 3.2.2 outlines the stages involved in the design of each corpus used in this research. The Romanian corpus, called RoTC, is specially built for this research[1] and its compilation stages with all the specifications regarding data collection, data preparation, and statistics, are reported. The Spanish corpus was made available for the current experiments[2], and its details are briefly presented. The section concludes with further details about the Weka toolkit, the machine learning software employed for the experiments described in the present thesis.

3.2.1 Background Concepts

The current subsection aims to define translational comparable corpora and to report the linguistic resources used in this research. To define this concept, other components need to be tackled beforehand, specifically those of corpus, and comparable corpus.

3.2.1.1 Defining a Corpus

Although there are several discussions on what constitutes a corpus (Sinclair, 1996; Leech, 1992; Francis, 1992; Atkins et al., 1992), it has been established that *a corpus is a collection of naturally occurring language data* (McEnery, 2003, p. 449), which is created according to the criteria

[1]Special thanks to Dr. Constantin Orăsan for the support provided with the non-translated component of the RoTC corpus.

[2]My sincere gratitude to my supervisor, Prof. G. Corpas, from University of Málaga, Spain, for making available the Spanish corpus for part of the experiments reported in the present thesis.

of assembling a body of *machine-readable authentic* texts *sampled* to be *representative* of particular language or language variety (Xiao, 2010*a*).

At the compilation stage of any corpus, general and particular aspects are considered depending on the purposes of the research, and obviously, the main characteristics observed by most scholars are those imposed by the definition itself, even though these allow for various interpretations and thus raise several questions and debates.

As expected, what can be considered as being *representative* is a widely discussed subject among scholars (Olohan, 2004; McEnery, 2003). This aspect is seen as the most important characteristic which differentiates a corpus from a "haphazard collections of textual material" (Leech, 1992, p. 116). It is difficult to ensure that the data is representative of a particular language or genre. When considering which texts should be included in the corpus, the decision-making process can go beyond text type or genre, text function or scope, how typical or influential the given text can be (Olohan, 2004, pp. 47–48). Also, regional and temporal factors can also be taken into consideration, being part of the criteria employed when building a corpus. Nationality, age, native language, ethnicity, etc., can all be decisive factors, depending on the research purpose.

Sample size is another relevant aspect which may be an important factor in achieving representativeness and refers to how many texts should be included in the corpus and what the ideal size of each of them should be. Representativeness depends on whether the sample includes the full range of language variability intended, so that the researchers using the corpus would be able to generalise their findings.

It is also emphasised that a bigger corpus is not necessarily more useful than a smaller one, as the amount of data under investigation is always limited (Kennedy, 1998, pp. 66-70). A smaller corpus can be sufficient in some cases, for example, if the research focuses on grammar only (Hunston, 2002, p. 26). The analysis of smaller corpora can also lead to remarkable discoveries: a large corpus is used for the discoveries of patterns in language data, whilst the smaller corpora are commonly employed in the comparison of different text types or genres (Sinclair, 2001). Ultimately, the factor accounting for the availability of suitable texts should not be dismissed in the creation of a corpus.

Salient issues arise in the process of deciding whether a new text should be included in the corpus, and the final choice can distort the data and compromise the research findings. Thus, corpus compilation is a recursive process that follows the principles agreed on in the theoretical analysis stage, and refines the final product until it meets the variation requirements imposed to achieve representativeness (Biber, 1993). It is not a truly completed action until the corpus is finalised and fulfils the entire set of conditions.

Making a connection to the descriptive translation studies area, it is important to focus on the following observation. Within this domain, it is pointed out that the effect of the source language on translations can make these types of text perceptibly different from native texts. As a consequence, it is suggested that translational language is "at best an unrepresentative special variant of the target language" (McEnery and Xiao, 2007*a*). For this reason, even in other domains, such as the natural language processing domain for systems like automatic summarisation or question

answering, this type of text is usually avoided in the compilation stage of corpora.

More specific characteristics need to be addressed in the compilation of a comparable corpus. The following paragraphs present the most frequently used definitions for this concept and emphasise the reasons why scholars need this type of linguistic resource in their research.

3.2.1.2 Defining a Comparable Corpus

Current research embraces the definition presented by McEnery (2003), who points out the key attributes of what constitutes a comparable corpus: two corpora, A and B, are considered to be comparable if both A and B are found to have:

- the same *sampling frame*[3] with *similar balance* and *representativeness*;
- the same *proportions* of the same *genres* in the same *domains*;
- the same *sampling period*.

Although many have attempted to define this concept as precisely as possible, scholars in the field have not yet reached agreement on a definition of comparable corpora. Nevertheless, there is a standard provided by EAGLES (1996), emphasising that a comparable corpus is *a corpus which comprises similar texts in more than one language or variety*.

[3]The sampling frame is an essential aspect of a comparable corpus. Both components involved in the corpus need to be matching with each other in terms of proportion, genre, domain and sampling period (McEnery and Xiao, 2007*b*, p. 133).

Considering the perspective shared by most researchers and the definitions discussed above, it appears to be a matter of how *similarity* can be understood or modelled depending on the research question. The degree of comparability is "in the eye of the beholder", strictly depending on the requirements and the objectives of the research study (Maia, 2003). Although several scholars have discussed this topic, the vagueness of the concept still persists, mainly because of the fuzzy notions used in its definition. Consequently, accurate illustrations of this notion are still deficient, and to what extent a corpus is comparable to another one remains a tricky question (Kilgarriff, 2001), from the starting point until the assessment stage.

Various views of the concept of comparable corpora manipulate uncertain terms, like: similarity, variety, domain, proportions, representativeness, balance. It is a rather complex task to draw a strict line between what is well balanced and what is not, or to define when a text is similar to another one, and, just as highly important, when this similarity is broken. The freedom offered by such a vague definition provides the opportunity to build several comparable corpora with varying degrees of comparability. The more flexible the degree of comparability, the more difficult it is to describe a valid, universal trend characteristic of the nature of the translated text.

Despite the controversy over the definition and the lack of agreed units to measure the degree of comparability of a comparable corpus, the need for these linguistic resources is unquestionable. Consequently, scholars need to adapt to existing circumstances, both practical (such as having to deal with copyright issues, with the lack of specific textual material, obtaining access

to information about the author or the source language of a text, etc.) and theoretical (such as assessing whether the selected texts are similar and representative enough in terms of *x*, *y*, *z* for the topic being explored). These circumstances often prove to be important factors affecting the entire compilation process.

Terminology Issues Besides the debates generated by attempts to define this type of resource, the fundamental notion itself has generated some misinterpretations and, as a result, inconsistency appeared in the terminology used by different scholars within the same research community.

For instance, a corpus with source texts and their translations can be seen as:

- a *translation corpus* (Grange, 1996, p. 38);
- a *parallel corpus* (Baker, 1993, p. 248), (Baker, 1995, 1999), (Hunston, 2002, p. 15).

At the same time, a monolingual corpus designed after the same sampling frame can be seen as:

- a *parallel corpus* (Grange, 1996, p. 38);
- a *comparable corpus* (Baker, 1993, p. 248), (Baker, 1995, 1999), (Hunston, 2002, p. 15).

Both types of corpus (the source texts and their translations, and the monolingual corpus compiled after the same sampling frame) are seen as

parallel by Johansson (1998, p. 4). However, many researchers adopted Baker's definitions of comparable and parallel corpora (Baker, 1993).

With a view to terminological consistency, the following aspects are considered in the classification of corpora:

- number of languages: a corpus can be a monolingual, bilingual, or multilingual corpus;
- content: a corpus can include translations or non-translations;
- form: a corpus can be either a parallel, or a comparable corpus.

To avoid further inconsistencies in the terminology, McEnery and Xiao (2007*b*) suggest that the criteria of content and form considered in the classification should not be mixed, and adopt the terminology described by Baker: *a parallel corpus is a corpus with source texts and their translations*, whilst *a comparable corpus is a corpus designed after the same sampling frame*. Comparable corpora can be *monolingual* (e.g., a comparable corpus of translated and non-translated type of texts written in the same language), *bilingual* (e.g., a comparable corpus of non-translated similar medical texts in two different languages) or *multilingual* (e.g., a comparable corpus of non-translated texts written in more than two languages).

Why compile comparable corpora?

Compiling comparable corpora for the investigation of various hypotheses proposed within the area of translation studies is currently one of the main tasks within the domain, and a time-consuming one. Nevertheless, the compilation of comparable corpora is required as it

appears to be the most suitable resource, given the research hypotheses investigated.

In translation studies, these hypotheses attempt to grasp and analyse certain features of translational language, hence the lack of resources proves to be a serious obstacle for the further refinement of ideas and findings, and consequently for the advancement of translation theory itself.

It is important to mention that this type of resource is not confined to being used in translation research only. It can also be used in other fields - for instance, for the improvement of statistical machine translation (SMT) systems. Scholars, such as Kurokawa et al. (2009); Lembersky et al. (2011), have shown that making use of the main hypotheses and findings of translation studies and training the SMT system on translational corpora can result in an overall improvement of their system.

The use of monolingual comparable corpora has been widely supported for the investigation of the nature of translational language (Baker, 1995; Corpas, 2008), and calls for the development of specific tools and resources for professional translators have had an impact on the domain. Even though few translational corpora have been built, such as the well-known English Translational Corpus (Laviosa-Braithwaite, 1996; Puurtinen, 2003*b*), most languages still lack proper resources for the investigation of translational hypotheses. To the best of the author's knowledge, Romanian is one of these languages.

The research presented in this thesis bridges this gap and reports on the compilation of the RoTC corpus, a monolingual comparable corpus based on newspaper articles. The RoTC corpus has been built as part

of the research reported in this thesis, and adheres to the requirements imposed by McEnery (2003) on comparable corpora. More details of the RoTC corpus can be found in Section 3.2.2.2.

As the nature of translational language is the focus of translation theory, compiling translational corpora is a vital resource underpinning the investigation of translational hypotheses. Several corpus-based approaches exploit comparable corpora, where comparability is obtained between translated and non-translated texts in the same language, as suggested by Baker (1995). Because the notion of comparable corpora has raised different interpretations among translation studies scholars, leading to the use of varying terminology, the following section aims to describe the notions of translational corpus and translational comparable corpus, respectively.

3.2.1.3 Defining a Translational Corpus and a Translational Comparable Corpus

This thesis considers that a corpus which comprises translated texts written by human translators is a *translational corpus*. Consequently, this type of resource is usually exploited within the area of translation studies in investigations into the nature of translated texts, but nevertheless, can be employed in other contexts.

For the investigation of translational hypotheses, the definition of a comparable corpus accepted as a standard, i.e., the one reported in EAGLES (1996), may allow for the appearance of a debatable issue: no translational corpus can be considered comparable since the resource only has texts in one language. Baker (1995) thus suggests that the concept of

translational corpus should be seen as a new type of comparable corpus. The resource proposed includes two subcorpora in one and the same language: one subcorpus with originally produced texts in a given language, and the other with texts translated into the same language from one or more source languages. Baker (1995) proposes that both subcorpora should be similar in terms of domain, variety of language, time span, and that they should be of comparable length.

For the investigation of hypotheses which compare assumed features of translated texts with those of non-translated texts, a corpus where comparability obtains between translated and non-translated texts in the same language can be considered an appropriate resource. In this thesis, this type of resource is considered a *translational comparable corpus*. If the translational hypothesis does not imply a comparison between translated and non-translated texts, then this thesis considers that a translational corpus, comprising only translated texts, may suffice.

In the investigation of translational hypotheses, namely, hypotheses which do imply a comparison between translated and non-translated texts, several corpus-based approaches make use of monolingual comparable corpora.

When studying language variation in translations, additional aspects arise at the corpus compilation stage. Apart from concerns about ensuring representativeness for a certain genre, the issue extends to whether the texts are representative enough to illustrate the behaviour of translational language (Olohan, 2004, p. 47), as this is the focal point of descriptive translation studies.

To provide an example, at the compilation stage of a parallel corpus, especially of a bidirectional one, the text selection stage may cause a loss of comparability (Zanettin, 2000, pp. 108-109). Zanettin (2000) explains this situation, emphasising the practical aspects of translated text availability, and the literary status of such texts: e.g., a corpus of English-to-Italian translations comprises mostly popular fiction, and Italian-to-English translations are more related to 'high culture'.

Also, the difficulties implied by the notion of *balance* constitute another important aspect, not only in terms of the type and status of texts, but also considering the reputation of the translators, or whether the texts chosen can be a definite criterion in the compilation of such a corpus. Crisafulli (2002) advocates text selection according to their literary value, an opinion based on an "idealised conception of literature"and translation (Crisafulli, 2002, pp. 32-33). To sum up, at the compilation stage, researchers face important decisions and the ideal resource is rarely actually built.

Nevertheless, despite the difficulties arising in the compilation process, there are linguistic resources available for the following widely-spoken languages: English (Baker, 1995), Portuguese (Frankenberg-Garcia, 2004), Spanish (Corpas, 2008), Chinese (Xiao et al., 2010), Italian (Baroni and Bernardini, 2006), Dutch and German (De Sutter and Van de Velde, 2008).

The Translational English Corpus, TEC, is probably one of the first corpora compiled for translation studies in the mid-nineties (Baker, 1995; Laviosa-Braithwaite, 1996), and the one which contains the most data on authors and translators. However, the entire corpus cannot be directly accessed due to copyright issues. The ten-million-word corpus comprises

four categories of texts: biography, fiction, newspaper texts and in-flight magazines, with translations into English from both European and non-European languages. The main experiments were done manually and they showed that corpus-based research allowed for translational hypotheses to be more clearly defined, for progressing to large-scale, target-oriented research, and for considering a wider range of social and cultural factors (Laviosa, 2002).

Besides the learning models reported in this thesis (see Chapter 4), this work reports on a new linguistic resource available for the study of translational language in Romanian. The next subsection provides details regarding the corpora used in the machine learning experiments conducted for the purposes of this study.

3.2.2 Translational Corpora Relevant to This Research

As translationese involves a comparison between the nature of translated versus that of non-translated language, the recommended linguistic resource for such an investigation consists of monolingual translational comparable corpora, containing translated and non-translated texts in the same language (Olohan, 2004). An approach based on this type of resource is more likely to avoid any foreign interference (Pym, 2008) and, consequently, it is more likely to fit in the investigation of the nature of translated versus non-translated language.

An observation on a specific issue regarding the compilation of translational corpora, as mentioned by Chesterman (2004*a*), is concerning

the representativeness of the sample selected for investigation. It is pointed out that the data may still be unrepresentative in some manner due to some neglected aspects, such as the so-called 'bad' translations: it is still unclear whether 'bad' translations should be included in the sample or not (Halverson, 1998). On this topic, Desmidt (2009) emphasises that, as the social context is relative and would generally lead to changes in translations and in the manner in which they are seen, then what actually constitutes a 'good' translation is also relative.

At this point, it is important to emphasise that the experiments conducted as part of this research do not have the objective to assess the quality of the translations investigated, hence the resource does not take into account this aspect. Moreover, even though the corpus would contain the so-called 'bad' translations, then the machine learning techniques would be able to handle real-life data, as this is their fundamental scope (Witten et al., 2011).

This research is conducted on two languages, Romanian and Spanish. The genres across these corpora are different: the Romanian corpus comprises newspaper articles, whilst the Spanish corpus comprises medical and technical texts. The main reasons behind this selection are the following: first, translation studies has largely investigated literary, religious, and philosophical texts, whilst most professional translators nowadays focus on more commercial, technical, and scientific domains (Olohan, 2007), and this resulted in a lack of attention to the genres widely used in the practice of translation to date; the second reason is more practical, as the texts from the first corpus are freely available for research

use, while the second corpus has been made available by the owner in order to be exploited in research experiments.

More details about the comparable corpora used are presented below, for each of the languages investigated: Spanish and Romanian.

3.2.2.1 Spanish Translational Corpus Description

The resource exploited in this work for the investigation of translationese and translational hypotheses for Spanish is the monolingual comparable corpus compiled by Corpas (2008). The entire compilation process of the corpus is described in Corpas (2008).

The resource comprises medical and technical texts produced between 2005 and 2008, written both by professionals and by undergraduate students in their final academic year at the Translation and Interpreting Department of the University of Málaga (Spain). The corpus has two subcorpora: one subcorpus contains translated texts and the other one non-translated texts. The corpus is structured in three pairs of translated and non-translated texts, as follows:

- Corpus of Medical Translations by Professionals (MTP), which is comparable to the Corpus of Original Medical texts by Professionals (MTPC);
- Corpus of Medical Translations by Students (MTS), which is comparable to the Corpus of Original Medical texts by Students (MTSC);

- Corpus of Technical Translations by Professionals (TT), which is comparable to the Corpus of Original Technical texts by Professionals (TTC).

Some of the components of the Spanish corpus are collected from the repositories of translation memory systems (TMs), having peninsular Spanish as target language, and some of them are collected ad-hoc according to imposed design criteria. The entire translated subcorpus (i.e., the MTP, the MTS, the TT) is translated from either American or British English into peninsular Spanish. Each pair is presented separately as follows:

- The MTP has the following characteristics: it comprises fragments of texts, not entire documents, which were collected from the repository of translation memory systems. It has biomedical texts varying from research papers published in journals to clinical essays, textbooks, product description and user instructions for surgical equipment.

  The MTPC, comparable to the MTP corpus, comprises non-translated biomedical texts and it has similar text types and topics as the MTP corpus.

- The MTS corpus consists of biomedical texts translated by final-year undergraduate students from the Translation and Interpreting Department of the University of Málaga (Spain). It comprises whole documents having nearly the same text types and topics as the MTP corpus, only with a higher proportion of research papers, product descriptions and patient information leaflets.

The non-translated comparable corpus to MTS, the MTSC, shares the same design criteria.

- The technical translated corpus, TT, comprises target language segments produced using translation memory systems. The texts belong to the technical and technological domains, covering topics such as telephony, network services, telecommunications etc. The subcorpus includes user manuals, guides and operating instructions, company press releases and, in a lower proportion, rules and regulations, standards, projects and monographs.

  The TTC corpus is compiled ad-hoc from evaluated electronic sources. Unlike the TT corpus, it contains whole documents, not only segments, and this may lead to a lower degree of comparability to the TT corpus. Moreover, it may prove to be an impediment in assessing the coherence features of the texts. Then, the texts were analysed in terms of text type, domain, topic, and selected to match the same design criteria as the TT corpus.

In conclusion, the Spanish corpus is a comparable corpus, as its translated and non-translated components meet the following requirements (Corpas, Mitkov, Afzal and Pekar, 2008):

- the pairs contain roughly the same text types and forms;
- they have texts in the same domains and sub-domains;
- they were produced within the same time-span: 2004 to 2008;
- they are approximately the same size in terms of number of tokens.

Spanish Corpus Statistics

Table 3.1 presents the values for the number of tokens for each subcorpus, and the percentage of texts that each subcorpus has in the entire comparable corpus. Note that there is a total number of 1,529,874 tokens for the overall translated subcorpus, and a total number of 2,696,079 tokens for the non-translated one.

Spanish Corpus			
Subcorpus	Tokens No.	Texts No.	Percentage
Non-Translated	2,696,079	294	65.33%
Translated	1,529,874	156	34.67%
Total	4,225,953	450	100%

Table 3.1: Spanish Corpus Statistics.

The ratio of 2:1 for the number of translated and non-translated texts is kept for the Spanish data as well, and the resulting values for the token numbers are fairly related to ensure comparability.

Spanish Corpus	
Subcorpus	Average
Non-translated	16,897.36
Translated	5,408.04

Table 3.2: Average Tokens per Document.

In Table 3.2, the average values for each subcorpus are illustrated. As described earlier, the components have texts from two distinct domains and they can be whole documents, or fragments of texts, or segments processed from the repositories of translation memory systems. As a result, the average values are influenced, and the non-translated subcorpus has a higher mean value than the translated component.

Note that the features extracted are all normalised at the text level since the machine learning models employed are classifying texts into translated and non-translated classes (see further details in Chapter 4 and 5).

In the current research experiments, a range of features needs to be extracted in order to investigate their impact on the nature of translational language. To extract these features, a parser for the Spanish language is required and, to this end, the Connexor Machinese (Tapanainen and Järvinen, 1997) has been used. Details regarding the output provided by the parser for the Spanish data are illustrated in the following paragraph, and Chapter 4 provides further details about the way in which the comparable Spanish corpus has been used in this research.

Corpus Pre-processing

The parser tokenises the text, creates links between the words and then names the links with corresponding syntactic relations, providing the following information: part of speech, morphological and syntactical information for each token, its lemma and dependencies. The tool can also provide its output in XML format, and a sample is represented in Figure 3.1.

In the current research, the XML output of the parser is then exploited to extract the required features under investigation. The features are assembled in a data type according to the input format imposed by the machine learning tool used for this work, namely Weka[4]. The

[4]http://www.cs.waikato.ac.nz/ml/weka

```
<sentence id="w1">
<token id="w2"><text>Integracion</text>
<lemma>integracion</lemma>
<tags><syntax>@NH</syntax>
<morpho>N FEM SG</morpho></tags></token>
<token id="w3"><text>de</text>
<lemma>de</lemma><depend head="w4">pm</depend>
<tags><syntax>@POSTMOD</syntax>
<morpho>PREP</morpho></tags></token>
<token id="w4"><text>tecnologias</text>
<lemma>tecnologia</lemma><depend head="w2">mod</depend>
<tags><syntax>@NH</syntax> <morpho>N FEM PL</morpho></tags></token>
<token id="w5"><text>de</text><lemma>de</lemma>
<depend head="w6">pm</depend><tags><syntax>@POSTMOD</syntax>
<morpho>PREP</morpho></tags></token>
... ... ...
</sentence>
```

Figure 3.1: The XML Sample from the Connexor Machinese Parser.

format required by this tool is described further in Section 3.3.2, after the fundamental theoretical concepts of machine learning are briefly introduced.

The details regarding the use of this corpus are reported in Chapter 4. The next section introduces a similar translational comparable corpora for Romanian.

3.2.2.2 Romanian Translational Corpus Compilation

The main objective in compiling the Romanian Translational Corpus is to allow for the investigation of translationese and related translational hypotheses, such as the well-known translation universals.

As no study of Romanian had been done for translationese, to the best of the author's knowledge, a dedicated type of resource did not exist. For this reason a comparable corpus has been specially compiled for this task,

consisting of newspaper articles published between 2005 and 2009. The RoTC corpus comprises two subcorpora, both pertaining to the journalistic domain: a translated subcorpus and a non-translated subcorpus. The translated texts are collected from the South-East European Times[5], a multilingual news portal translated into nine Balkan languages, Romanian included. The translated subcorpus comprises 223 articles written between 2005 and 2009 to preserve the same time-frame characterising the non-translated subcorpus. The non-translated subcorpus comprises 416 texts, from a well-known Romanian newspaper called 'Ziua'[6].

Collection and Selection of Data

The content of the "South-East European Times" website is in the public domain, meaning it can be used and distributed without permission. The process of selecting the articles for the RoTC corpus is described in the following paragraphs.

All the articles were downloaded using various scripts which use the URL structure information. The link allows the selection of the articles to suit various needs; in the given context, these are:

- selecting articles according to their language (i.e., the URL contains the string "www.setimes.com/ .../ro/... "for the Romanian language);
- selecting articles according to the date (i.e., the date can be easily extracted from the link as it appears in this format "www.setimes.com/ .../yyyy /mm/dd/... ").

[5] `http://www.setimes.com`

[6] `http://www.ziuaveche.ro`. The name of the newspaper can be translated into English as 'Daytime'.

The topic of the selected articles was international news covering roughly the same subjects over the same time-span, so as to obtain a comparable corpus consisting of texts selected from the 'South-East European Times' website and the 'Ziua' newspaper. Also, the number of texts has been balanced by randomly selecting 416 non-translations written between 2005-2007 versus 224 translations written between 2005-2009. Because it is important to have balanced data in a study which adopts the machine learning approach, in this case, between translated and non-translated number of texts, a ratio of 2:1 is maintained for this resource.

Corpus Composition

The RoTC corpus has a total of 341,320 tokens (200,211 for the translated subcorpus and 141,109 for the non-translated subcorpus). It is very likely that the selected articles are written by various translators, so the possibility of a specific style playing a role in the classification task is avoided. It is also extremely likely that the texts are translated from various languages into Romanian, an advantage that ensures a high likelihood that all the patterns discovered are not due to one particular source language. The shortcoming of the translated subcorpus is that the portal, due to confidentiality issues, fails to provide precise information about the source language or the identity of the original author or the translator.

Due to resource scarcity for the Romanian language, and in order to ensure the comparability that exists of the two sub-corpora, all articles selected from the portal were considered to be translations, even if there is a chance of some articles being originally written in Romanian. Nevertheless,

some of the articles mention the source of their information (e.g., Reuters) and the original source language of the given text can thus be deducted. In addition, it is often stated that various information sources were used when the given article was produced.

The argument that the articles are translations and not original, non-translated texts is inferred from two distinct sources: Firstly, it is inferred from the following rationale: one text cannot be originally produced in ten languages and yet be perfectly aligned from one language to another (i.e., one Romanian article to have its source language Romanian, the corresponding, parallel Turkish article to have its source language Turkish, and at the same time, both the Romanian and the Turkish news to be perfectly aligned to each other).

The fact that all the articles are aligned to each other leads to the assumption that at least nine out of ten parallel texts are in fact translations. Consequently, it is highly probable to have mostly translations, if not only translations, in the RoTC translated subcorpus. However, the attempt to clarify this aspect from its source failed due to the portal's confidentiality policy.

Secondly, the assumption that the portal comprises translated texts is inferred from the following source: the portal was also entirely harvested and used in a machine translation task, reporting the resource as having translations into Balkan languages, including Romanian (Tyers and Alperen, 2010).

The non-translated subcorpus does not present the same difficulty in assessing whether the texts are non-translations, since 'Ziua' is a national

newspaper with news written only in Romanian and dedicated to local nationals. Moreover, the articles do state their authors, and their full names indicate that they are Romanian natives. It is thus concluded that the subcorpus comprises non-translated texts, written by various authors.

Corpus Pre-processing

In the pre-processing stage of the compilation of the corpus, all the texts were tagged using the part of speech tagger provided as a web service by the Research Institute for Artificial Intelligence[7], part of the Romanian Academy (Tufiş, Ion, Ceauşu and Ştefănescu, 2008; Tufiş, Ştefănescu, Ion and Ceauşu, 2008), and its output transformed into XML[8] format to facilitate access to the data representation of the document[9]. A sample of the XML format is represented in Figure 3.2. A few statistics about the size of the RoTC corpus and its components are reported below.

RoTC Corpus Statistics

Some fundamental statistics are computed for the RoTC corpus. In Table 3.3, the size of the corpus is presented as the number of tokens for each subcorpus, and as a whole. Note that the RoTC corpus has a slight majority of non-translated texts, accounting for 58.66% of the total number of articles.

This happened as the amount of texts available for the same topic in the comparable translated corpus is slightly lower compared to the number of non-translated articles, and the intention was to obtain as many articles

[7] `http://www.racai.ro/webservices/`
[8] Extensible Markup Language
[9] See details about data representation in Chapter 4, Section 3.3.1.2.

```
<sentence id="w128">
<token id="w129"><text>Acestea</text>
<lemma>acesta</lemma><tags>
<morpho>Pd3fpr</morpho></tags>
</token>

<token id="w130"><text>au</text>
<lemma>avea</lemma><tags>
<morpho>Va--3p</morpho></tags>
</token>

<token id="w131"><text>fost</text>
<lemma>fi</lemma><tags>
<morpho>Vmp--sm</morpho></tags></token>

<token id="w132"><text>primele</text>
<lemma>prim</lemma><tags>
<morpho>Mofprly</morpho></tags></token>

<token id="w133"><text>alegeri</text>
<lemma>alegere</lemma><tags>
<morpho>Ncfp-n</morpho></tags></token>
... ... ...
</sentence>
```

Figure 3.2: Sample of the Output Provided from the POS Tagger Converted into XML format.

as possible so as to be able to use the resource in a machine learning framework. Obviously, comparability aspects were considered, so it was decided to maintain a ratio of 2:1 between the translated and non-translated texts in order to comply also with the same sampling frame with a similar balance factor.

Table 3.4 reports on the average value in terms of number of tokens per text. The figures show that the RoTC corpus has an average number of tokens of 481 for the translated subcorpus, and 632 for the non-translated texts. These values are closely related as expected since there are only newspapers articles in this corpus. It remains to be further investigated

RoTC Corpus			
Subcorpus	Tokens No.	Texts No.	Percentage
Non-Translated	200,211	223	58.66 %
Translated	141,109	416	41.34 %
Total	341,320	639	100%

Table 3.3: RoTC Corpus Statistics.

whether this slight difference is due to some feature assumed to be specific to either translational language or to non-translational language (some hypotheses make reference to the size of translated texts in general). Nevertheless, the RoTC corpus also complies with the same proportion requirement for a comparable corpus.

RoTC Corpus	
Subcorpus	Average
Non-translated	632.78
Translated	481.28

Table 3.4: Average Tokens per Document.

For the Romanian learning model proposed in the current research, the training set comprises 639 randomly selected news articles and the overall test set has 148 randomly selected articles. The same text ratio, 2:1, is kept for both selected training and test datasets of the learning models employed within this work. This is a restriction imposed by the approach adopted, machine learning, to be able to have a balanced number of instances in the training data (see details in Chapter 4).

The fundamental concepts about machine learning domain are briefly pointed out in the next section.

3.3 Machine Learning with Weka

The discipline of machine learning is located within the data mining area. Vast amounts of electronic data are becoming more readily available and, as a result, it becomes more and more necessary to be able to harvest efficiently the valuable information they contain. Data mining is the discipline which analyses data and uses software techniques to find patterns and regularities within sets of data. The nature of data mining is thus interdisciplinary, involving the following academic fields: databases, statistics, machine learning, computer science, visualisation, mathematics (Mitchell, 2006).

3.3.1 Preliminary Notions about Machine Learning

Being scarcely adopted within translation studies research, the massive discipline of machine learning is briefly introduced in this subsection, together with its main notions.

3.3.1.1 What is Machine Learning?

Machine learning is a subfield of Artificial Intelligence concerned with the design and development of algorithms and techniques that allow computers to "learn".

The discipline of machine learning has emerged at the confluence of Computer Science and Statistics, and it thus combines the main questions of them both, giving birth to a new central one. Whilst the question that computer science aims to answer is *"how can we build machines that solve*

problems?", and the question that defines statistics is *"what can be inferred from data and a set of assumptions, and with what reliability?"*, the machine learning discipline is influenced by both of them, and deals with a different question: *"how can we build computer systems that automatically improve with experience?"*(Mitchell, 2006).

Defining the concept of automatic learning, Mitchell (1997, p. 2) states that a computer program is said to *learn* from experience E with respect to some set of tasks T, with a performance measure P, if its performance at the tasks in T improves with the experience E.

A classic example of such a program is the automatic detection of spam emails. The task T is the spam filter, P would be the percentage of correctly identified spam emails[10], and the experience E is the set of emails which were labelled as being 'spam' or 'not-spam'.

A further question is *"when is it suitable to adopt a machine learning approach?"* and an answer would be: whenever the aim requires discovery of knowledge across hidden regularities in large data sets. Obviously, this can imply a broad range of suitable situations in which machine learning proves to be useful, or at least is expected to be.

As an example, thinking of various needs that language imposes, machine learning is frequently adopted in the domain of natural language processing. It proved to be an excellent approach for many applications, such as: syntactic pattern recognition, search engines, speech and handwriting recognition, machine translation. For instance, given a set of letters handwritten by several people in different ways, the learning system

[10]P is also known as the accuracy of the system.

has the task to correctly identify a particular letter from all the possible letters.

Circling back to the research gap pointed out in translational hypotheses research (i.e., the quest for patterns characteristic to translational language), machine learning appears to have a remarkable potential in discovering hidden regularities in large data sets representative of translational language. By bridging this gap, a significant advance within translation research may be noted, which can also pave the way for refining the existing theories and beliefs about the nature of translational language.

Next, the machine learning concepts relevant for this research are briefly illustrated in the following paragraphs.

3.3.1.2 Main Concepts

It can be overwhelming to assess a large set of features extracted from vast amounts of texts in order to test if they support a certain translational hypothesis. For this reason, machine learning techniques give the opportunity to simultaneously analyse the set of characteristics proposed for a certain hypothesis, in this case, for translationese or any translational hypothesis under investigation.

Learning is acquired from examples, also referred to as *instances*, and an example is described by a set of *features*, also known as attributes or characteristics (a few examples follow). The array containing the value for each feature of an instance is the *feature vector* used in the learning system. In the machine learning domain and throughout the following

chapters, the following terms are used interchangeably: 'feature vector' or 'data representation'.

By convention, the last feature appearing in the data representation is the *class* of the learning model, representing the target concept for learning (i.e., the concept which the model tries to learn). For instance, a common task in the machine learning domain is the classification task: the model has the aim to categorise instances into a set of classes (i.e., the task to assign an example to one of the categories provided as values of the class attribute).

For instance, given a set of attributes, such as 'outlook', 'humidity' and 'wind', a learning model has the task to assess 'whether John will go to play tennis or not'. More information on the attributes follows: Outlook has three values: 'sunny', 'overcast', and 'rain'. Humidity has two possible values: 'high' and 'normal'. Wind has the potential values: 'strong' and 'weak'. This is just a basic example, obviously the attributes can be more refined and not necessarily presented as nominal features[11]. The learning model attempts to correlate these features to the class given, namely 'John goes to play tennis' or 'John does not go to play tennis'.

Another important aspect in any machine learning model is the type of learning, which can be classified as *supervised* or *unsupervised*, depending on how the training data (experience) is represented.

If the training data has the examples annotated with their corresponding categories marked in the class attribute (labelled instances), then it is supervised learning. Otherwise, the learning is unsupervised. For

[11]See Chapter 3, Section 3.3 for types of attributes.

instance, the set of emails which have the label 'spam' or 'not spam' falls into supervised learning, whilst a group of unlabelled emails would require an unsupervised learning approach. As an observation, in this research the learning model is a classification task using a supervised learning approach.

At this point, the concept of *classifier* is introduced. The algorithms or techniques used in the learning model are known as *classifiers*, or *learners*. In supervised learning, the classifiers are presented with training examples that show the relation between input and output values. In unsupervised learning, the classifiers are expected to approximate the correct output, i.e., the class.

Weka toolkit is relevant to this work as it provides several classifiers for machine learning investigations (Bouckaert et al., 2012). It is largely used across different research communities, making an outstanding contribution to the domain and becoming a landmark system in the history of data mining and machine learning.

3.3.2 Data Preparation for Weka

In the pre-processing stage, the range of attributes whose selection process is described in detail in Chapter 4 is extracted from the corpus under investigation, and their values are mapped onto the format required as input for the Weka system, in this case, the ARFF format[12]. This format is presented according to the manual which accompanies the machine learning tool (Bouckaert et al., 2012).

[12]ARFF is an acronym for Attribute-Relation File Format

Dataset is one of the basic concepts of machine learning and it comprises, as the name suggests, the set of data items fed into the machine learning software (Bouckaert et al., 2012). A dataset can also be seen as a two-dimensional database table, comprising a collection of examples, each of them called an *instance*. These examples are required for a classifier to learn how to predict the corresponding class attribute.

Another fundamental concept for any learning model is its *evaluation*. In order to evaluate how accurately the classifier is able to predict the class of the model (i.e., referred to as classifier's performance or accuracy), there are different methods of evaluation. All the methods use a *training dataset*, for the training process of a classifier as the name suggests, and a distinct *test dataset*. The test dataset has instances that were not seen in the training process of the model.

The evaluation methods relevant to this research are the following: the 10-fold cross-validation evaluation, and the test dataset evaluation. The former uses nine parts, also named *folds*, of the entire training dataset and it evaluates the performance of the classifier on the tenth part, whereas the latter method uses the entire training dataset for the training process and it evaluates the performance on a separate test dataset.

Each instance consists of a list of attributes, also named features, which can fall into three categories:

- nominal, when the attribute has a predefined list of values: e.g., it is known that a corpus can be either monolingual, or multilingual; this information could constitute a nominal attribute with two known values: monolingual, multilingual.

- numeric, when the attribute is a real number or an integer: e.g., the proportion of nouns in a text is a real number which results in a numeric type attribute.
- string, a list of characters: e.g., the file name of each document is a string type attribute.

Figure 3.3 illustrates a sample of the dataset prepared for the Weka input, and its components are explained in the next paragraphs. In the structure of any dataset, the keyword *@data* is marking the part where all the instances are located. In this sample, there are only two examples in the *@data* due to space restrictions (i.e., one instance for the *translated* class and one for the *non-translated* class).

```
@relation 'all-features-ro'

@attribute fileName STRING
@attribute GrammaticalWords real
@attribute Nouns real
... ... ...
@attribute SentenceLength real
@attribute WordLength real
@attribute SimpleSentences real
@attribute class {"translated", "non-translated"}

@data
"text-1.xml", 0.26,0.29,[..],29.26,5.26,0.73,"translated"
"text-147.xml", 0.28,0.27,[..],28.74,5.35,0.68,"non-translated"
... ... ...
```

Figure 3.3: The ARFF Sample Format.

A dataset has the following structure: first, an internal name of the dataset is stated after the keyword *@relation*, here the dataset is called "all-features-ro", then the next lines define the attributes and their types using the keyword *@attribute*, and the instances included in the dataset

appear after the keyword *@data*. The *@data* part has comma-separated values for each of the attributes previously defined, in the same order as the features were defined. Each line of the *@data* represents one *instance* of the dataset, or sometimes referred to as *an example*. In other words, the list of the attributes preceding the *@data* represents the headers of the table's columns, and each line from the *@data* is the corresponding row.

The last attribute, always found as *@attribute class*, enumerates the possible values of the categories for the classification task. In this case, the class values are "translated" and "non-translated"for each instance of the dataset, since the objective is to distinguish between these two categories.

The class values appear at the end of the attribute values for each instance in the dataset, and they are used for training a classifier, in order to learn associations between attributes and classes. In the evaluation of a learning model, the existing class values are ignored while the classifier attempts to predict the class values according to the experience it acquired in the training process. The actual class values are used only to evaluate the accuracy obtained by that classifier.

Once the training and test datasets are mapped into the required *.arff* format, these can be processed with the machine learning techniques provided by Weka.

3.4 Conclusions

This chapter introduces the resources used for the investigations undertaken in this thesis on the nature of translated texts. These resources include

monolingual comparable corpora: one for Romanian, the RoTC corpus, specially compiled for the experiments included in this research, and one for Spanish, and tools necessary in the pre-processing stage of the corpora as well as the machine learning tool relevant to this research. Both corpora comprise translated and non-translated texts following the comparability requirements of such a resource and are described in detail in Section 3.2. To prepare the data for the extraction of the features analysed in this research, specific natural language processing tools are necessary: a part-of-speech tagger was employed for Romanian, whereas a dependency parser was used for Spanish. Both these tools are described in Section 3.2.2.

As the present research adopts the use of machine learning techniques, the fundamental concepts pertaining to this discipline are briefly introduced in Section 3.3, along with the machine learning tool relevant to this research, Weka. The learning model requires a set of potential characteristics of translational or non-translational language, features which are automatically extracted using natural language processing tools.

The next chapter of this thesis provides the justification for the need for machine learning approach in Descriptive Translation Studies, and reports on the learning models created for the investigation of translational hypotheses.

Chapter 4

Investigation of Translational Language from A Machine Learning Perspective

4.1 Overview

After having established the context of this research by describing translationese and related translational hypotheses in Chapter 2, and by presenting the resources relevant to the current research in Chapter 3, this chapter reports on the methodology adopted in this work for investigating the nature of translational language.

This research is closely related to three areas: translationese and translational hypotheses research; machine learning; and natural language processing. The theoretical concepts come from the first domain, whereas the methodologies pertain to the last two areas. These three areas are

relevant for the following reasons: first, the overall goal of the research is to investigate to what extent can hypothesised features of translational language distinguish translated texts from non-translated ones, a research question which places this work principally in the domain of translation studies; second, the method chosen to analyse them is the machine learning approach; third, the approach adopted requires a set of indicators which are automatically retrieved using natural language processing tools. Details about the interdisciplinary nature of this research are provided in Section 4.2.4.

The machine-learning approach is modelled as a text categorisation task in the investigation of the translationese hypothesis, aiming at learning to distinguish between translated and non-translated texts. A set of characteristics which are hypothesised to distinguish the two types of text is employed in the learning model, which is referred to as the translationese learning model. Additionally, the learning model also uses features motivated by the simplification and explicitation hypotheses to analyse to what extent they influence the translationese learning model.

This chapter is divided into three main sections: Section 4.2 emphasises the need for the machine learning approach and the natural language processing tools in the investigation of translational hypotheses, whilst Section 4.3 illustrates the translationese generic learning model, the simplification learning model, and the explicitation learning model for both languages: Spanish and Romanian. An overview of the entire learning framework is reported in Section 4.3.1, emphasising distinct sets of experiments, named *research scenarios*. Another brief section, more

precisely Section 4.4, is pointing out the assumptions considered throughout the learning models.

The chapter is summarised in Section 4.5, and the findings for each research scenario are reported and analysed in Chapter 5.

4.2 Direction of Research and its Benefits for Descriptive Translation Studies

The goal of this research is to investigate a considerable amount of translational data in a rigorous manner, adopting a computational approach, in order to retrieve indicative patterns of translational language[1]. The reasons for adapting a different direction of research in this thesis are highlighted in the following section, while Section 4.2.2 introduces the main strengths and core concepts of the machine learning based framework underlying the present work.

4.2.1 The Need for a Different Approach

Embracing corpus-based techniques in the quest for validation of the translational hypotheses has had a well-known impact in the domain, leading to several advances pointed out in Chapter 2. However, a considerable amount of investigations are conducted manually or semi-automatically, using rather small resources, factors which may lead to statistically insignificant results, especially for hypotheses suggested to

[1]Details on the definition of translational patterns are provided in Section 4.3.

be universals. Most of the studies use translations from one language to another one, aiming at observing patterns that occur in translational language in general although the study is only analysing a small amount of data (Gentzler, 1993; Frankenberg-Garcia, 2004).

Certain drawbacks appear in these investigations, such as:

- the methodology employed does not provide a rigorous evidence towards the validity of translational hypotheses, and the results yielded are often difficult to interpret (Laviosa-Braithwaite, 1996). Only a few studies adopt other methodologies, such as Baroni and Bernardini (2006); De Sutter et al. (2012), providing distinct views on translational language. The corpus-based approach is insufficient for hypotheses which aim at generalising certain tendencies in the language. Quantitative studies with statistical algorithms on large amounts of data are appropriate for this purpose. Most likely the features investigated in these quantitative studies need to be extracted from the observations drawn in the qualitative analysis of translational language.

- due to the hard-labour aspect of manual work, only a few studies investigate more than one hypothesis at a time, even though the literature suggests that there are a couple of potential hypotheses regarding translational language. This aspect may pose difficulties in explaining and interpreting the results obtained.

- due to limitations of the manual analysis, a low number of features are investigated within each study. This leads to a limited perspective on the nature of translated texts since many more features influence

the language of translations, all of them at once, and there may even be unexpected correlations among them.

- the inability to compare the results yielded on distinct studies aiming to validate translational hypotheses. This shortcoming appears because of the lack of a common multilingual methodology within the domain (Mauranen, 2008). To provide an example, Laviosa-Braithwaite (1997) analyses overall word frequencies, whereas Jantunen (2001) focuses only on a limited set of individual items. Both of the studies investigate the simplification hypothesis, although their outcomes cannot be compared. Consequently, the universality factor widely discussed among scholars does not have clear means of investigation.

- the lack of ranking among the suggested features of translational language. This thesis points out that there is a gap in terms of ranking among the features proposed according to which extent the characteristic appears to make a considerable distinction between translated and non-translated texts. This type of ranking is necessary in order to refine the current beliefs and assumptions regarding the nature of translational language, and to further advance the theoretical background of the discipline. To the best of the author's knowledge, such research study that provides a ranking of the features which characterise translational language does not exist.

- translational hypotheses still lack rigorous evidence to support their claims (Becher, 2011*a*).

Considering the above shortcomings, a suggestion for improvement would be a multilingual methodology, able to handle a larger set of features at a time for the same data, providing a ranking across the features explored. This would enhance the perspective over the phenomena occurring in translational language.

4.2.2 Taking the Machine Learning Turn in Descriptive Translation Studies

The approach proposed in this thesis fits the above requirements, and is similar to those typically used in data mining as it relies on machine learning techniques. This approach is in line with the following points emphasised in the literature:

- to design a methodology that provides computational power and uses statistical algorithms to test, investigate and identify potential features of translational language (Baker, 1993, p. 243);
- to retrieve patterns that occur in the translational language in order to provide the grounds to understand the phenomena occurring in translation. This would represent a major advancement in the domain (Chesterman, 2004*a*, p. 11).

The field fails to provide a methodological framework able to automatically search for such patterns, and at the same time, a methodology applicable to different languages. A multilingual methodology would pave the way towards the discovery of universal features and general

patterns likely to occur in translational language, irrespective of the source or target languages involved.

Although machine learning is rarely employed in the investigations on translational language, this type of technique can have an important impact. First, if the machine learning techniques are able to identify translated from non-translated texts based on the assumptions/features indicated within the literature, then two points have rigorous evidence: namely, the translationese and the features which proved to be relevant in the classification task. Second, the machine learning domain has learning algorithms able to reveal patterns used in the classification task, indicating the features they relied on in the learning process. Thus, they have the potential to reveal important patterns of language which occurs within translated texts, paving the way to more rigorous hypotheses regarding the nature of translated texts.

Consequently, a change in the research methodology is advocated in this thesis. This work advocates a machine learning approach in conjunction with the use of natural language processing tools. The shift is in line with the current computational trends within translation studies.

This research investigates the nature of translated texts by modelling a learning framework towards the automatic categorisation of translated and non-translated texts. The above issues are addressed, and the thesis reports a learning model able to analyse and rank different features of translational language, providing a set of patterns extracted in the learning process. The description of the framework is presented in Section 4.3.

The potential of the machine learning approach in an investigation of translational hypotheses is next highlighted. As machine learning systems are rarely adopted in translation studies, their main strengths and the core concepts of such a framework are briefly introduced in the following section.

Machine learning is one approach that can be considered in the investigation of the different features proposed as characteristic of translational language, and consequently, in the quest for patterns characterising translational language. A natural question arises when any potential methodological turn is suggested in a domain, such as: *"why adopt machine learning in the investigation of translationese or any translational hypotheses?"*. To answer this question, a few strengths of this field need to be outlined.

Machine learning being applied across several domains (e.g., natural language processing, computer vision, cognitive science, biology, etc.), is chosen for the following reasons:

- Machine learning is sufficiently *flexible* to be applied in various contexts. Due to its fundamental statistical-computational theories of learning processes, machine learning can be easily applied across diverse domains (Mitchell, 2006; Witten et al., 2011), and it has registered outstanding results for tasks such as: medical diagnosis, bioinformatics, detecting credit card fraud, stock market analysis, game playing, computer vision and robot locomotion. Thus, the machine learning approach can be modelled towards the investigation of several hypotheses proposed within the translation studies domain.

- Machine learning is known for its power to *efficiently mine data in the search for patterns*, a necessary requirement if the hypotheses under investigation are set to discover patterns from translational language. This aspect is almost self-explanatory as to why is it important for translationese research: it would extract patterns that arise during the statistical analysis of the training data[2] at hand.

- The ability *to analyse a large range of features at the same time* is an important strength of this approach. As the number of features increases, it becomes more difficult to assess them simultaneously, or to make statistical correlations among them, or to associate certain features with a trend in translational language. Considering this aspect, the machine learning approach would bridge a major gap in the literature: investigating a large set of features at the same time, being able to make statistical correlations between them, and even ranking the features according to the learning task.

- Most or all machine learning algorithms *can handle noisy data* (i.e., bad training examples). The features proposed to support certain hypotheses can be employed in a machine learning system, regardless of whether the hypotheses overlap or even contradict themselves in the literature. Machine learning techniques are statistical algorithms created to consider real-life data, which are frequently noisy. The algorithms are able to discard the uninformative features from the data representation, selecting only a subset of them, according to their influence on the learning task. This characteristic is particularly

[2]See Section 3.3.1 for more details on training data.

useful for the present thesis since this research uses natural language processing tools which are likely to introduce a degree of noise.

To sum-up, the machine learning approach appears to be a promising method in the domain. Besides the advances given by the adoption of corpus techniques in the investigations on the nature of translated texts, the machine learning techniques can discover patterns of translational language, and can point out correlations among the features analysed in texts, and even rank them according to their influence in the learning task.

How the features used in the learning approach are automatically extracted from large amounts of text is explained below, in the next section.

4.2.3 The Need for Natural Language Processing Tools in Descriptive Translation Studies

The potential interest that natural language processing tools may hold for descriptive translation studies can be summed-up as follows: to be able to harvest and analyse certain linguistic features on large textual material, the adoption of automatic tools to process natural language is required. This domain is entirely dedicated to creating these types of tools or even entire frameworks having various aims (e.g., automatic annotation of temporal relations and expressions in texts (Marşic, 2011), resolving anaphora (Mitkov, 2002) or distinct types of anaphor (Mihăilă et al., 2011), etc.).

These tools can help extract different features from texts, and then, the characteristics can be further analysed using machine learning techniques. To give an example, these three domains can interact as follows:

- translation studies sets the major goal; in the present thesis that is to find features and patterns that are able to automatically make the distinction between translated and non-translated texts;
- natural language processing domain provides the tools necessary to extract a set of features to be investigated;
- machine learning domain provides the techniques to learn to differentiate between translated and non-translated texts given the set of features provided.

The current research uses these tools and techniques to retrieve translational patterns and to create a multilingual model able to differentiate between translated and non-translated texts. The theoretical concepts pertain to translation studies, but the methodologies belong to machine learning and natural language processing. In Section 4.3, a translationese learning framework is reported, describing the attributes used and why they were chosen to characterise the target concept.

4.2.4 Interdisciplinary Study

Adopting the machine learning approach, the viewpoint of this thesis develops from the intersection of two disciplines: descriptive translation studies and machine learning. The defining question for descriptive

translation studies is *How can we describe the nature of translational language?*, whilst for machine learning it is: *Can computers learn from experience? And with what reliability?* Then the emerging question becomes *How can computers learn to discover hidden regularities or patterns of translational language, and with what reliability?*

The third area, natural language processing, is involved in the automatic extraction of the features to be mined by the learning model in order to retrieve potential patterns. As the data to be analysed is textual data, another question is implied: *how to mine textual data (corpora) for the discovery of patterns in the nature of translational language?* The investigation would thus require the use of natural language processing tools for the extraction of the sought features from the textual data.

The natural language processing domain already has relevant applications built for different tasks, and may easily provide the necessary tools for the automatic retrieval of attributes to investigate in the translational language, e.g., sentence type, or more complex, the identification of a certain type of ellipsis in texts. From the natural language processing perspective, this thesis adopts the use of parsers and part of speech taggers, tools which are described in Chapter 3.

Considering all these points, it can be inferred that a learning model can be built to retrieve patterns of translational language using a learning system designed as a categorisation task, aiming at distinguishing translations from non-translations. Moreover, since the corpora used for this research have the texts annotated as translated or non-translated, the

type of machine learning approach chosen for this task is the supervised one.

4.3 The Learning Models

The main objective of this thesis is to discover patterns of translational language. To this end, several potential features, hypothesised to be specific to translationese, are aggregated in a learning model[3].

At this point, it is important to emphasise what a pattern in the context of this research means. A *translational pattern*, or a pattern of translation, is a correlation between one or more quantifiable characteristics and the translational language. Once a pattern appears to reliably identify translated texts, its characteristics are then seen as *translational features*. A small note on the fact that the translational features are not automatically assumed to be universal (specific investigations are required to this end).

For example, let us assume there is a feature F that has a clear formula to be calculated and a monolingual comparable corpus which comprises translated and non-translated texts. May this feature F be the proportion of anaphoric pronouns in a text. Calculating F for each text, a pattern retrieved by a system can hypothetically appear as follows: "*if* $F \leq 0.035$ *then the text is a translation*". Note that the manner in which a pattern is represented can differ, but nonetheless, all of them emphasise a correlation between a set of quantifiable features and the concept sought (i.e., the translated texts given in the example provided).

[3]Note that fundamental concepts about machine learning are presented in Chapter 3.

The learning model has the task to automatically distinguish between translations and non-translations. Additionally, the model adopts features previously proposed in the literature in its investigations to be able to analyse related translational hypotheses, namely simplification and explicitation.

This research is based on the following rationale: assuming that translationese exists, then this research presupposes that translational language does have its own, specific, typical features through which humans can distinguish it from the non-translational language. Thus, an automatic system able to separate translated from non-translated texts can be designed, considering the potential features of translationese outlined in the literature.

In order to automatically retrieve a set of potential characteristics of translational language, and then aggregate them in a machine learning model, three distinct domains are combined for this research. The next section outlines how these domains interact for the objective sought in this research.

4.3.1 Structure of a Learning Model

Using the notation presented in Chapter 3, Section 3.3, the learning model can be described as:

> The task T aims at identifying translated and non-translated text. The experience E can be acquired from the labelled comparable corpora since the comparability is between

> translated and non-translated texts, and the resource has this information marked[4]. The learning process thus needs to analyse and correlate the features from the data representation of the model, and employ the use of learning techniques for the categorisation task. The system's accuracy P reflects how reliably the classifiers are able to distinguish between translations and non-translations.

Research within the machine learning domain has developed techniques for the classification task, and some of the best known are employed in this research. In the next sections of this chapter, the feature vectors proposed for each learning model are reported.

Translationese, simplification and explicitation can be studied by comparing translations with non-translations in the same language (Olohan, 2004), thus strictly avoiding any foreign interference (Pym, 2008). The main resource to be used in the investigation is the monolingual comparable corpus composed of translated text vs. comparable non-translated text[5].

Although the universality aspect implied by the simplification and explicitation hypotheses is not the subject of investigation in this research, the approach adopted provides a portable and easily adaptable framework which can ease further investigations on distinct languages testing the same, or almost the same, set of features. This is obtained by the use of multilingual features in the data representation. Consequently, the features can be analysed whether they appear to be reliable for any other

[4]See the description of the Spanish and Romanian corpora in Chapter 3.
[5]The Spanish and Romanian corpora are presented in Chapter 3.

target language, and if so, their universality may be thus tested for other translational data.

In Figure 4.1, an overview of the main processing stages involved in the development of the learning models used in this research is presented.

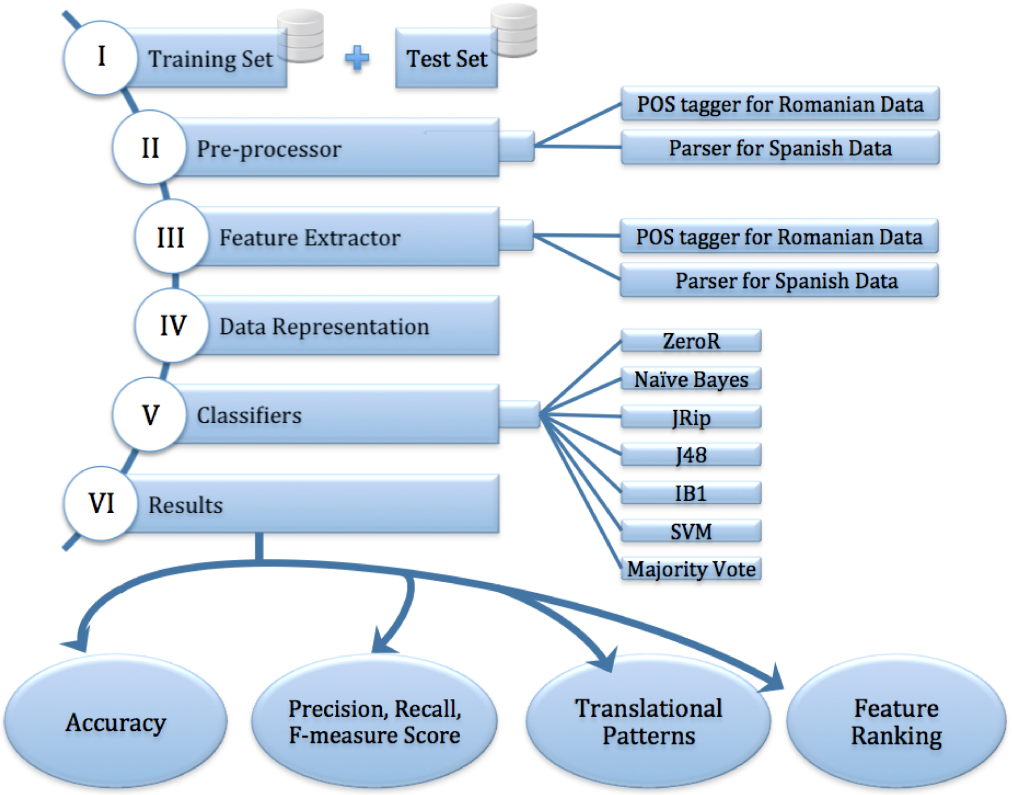


Figure 4.1: A Learning Model Overview.

In the diagram, each stage of the framework is illustrated. First, a training dataset and a test dataset are built comprising random instances from both classes. The same ratio of 2:1 between translated and non-translated instances, and also between the medical and technical texts used for the Spanish model, is preserved to ensure a balance in the learning process. Otherwise, the system would tend to learn better only one of the classes.

Second, the data is pre-processed. Given that currently there is no parser available for Romanian, the Romanian data is processed with a part-of-speech tagger, whereas the Spanish one with a dependency parser. A slight difference between the set of features extracted for these languages appears as a consequence of the natural language processing resources available for each of them. These differences are pointed out in the discussion of the features proposed for the learning models in Section 4.3.2.

Obtaining the output from these tools[6], the third phase extracts the required set of features for the two languages, Spanish and Romanian.Then these features are structured in the expected format for the machine learning application used, namely Weka[7], forming data representation for the learning model. Each instance represents one text, belonging either to the translated text class, or to the non-translated text class.

The fifth phase is the machine learning stage using a set of typical algorithms. As the target concept is translationese, a hypothesis which assumes that translated and non-translated texts differ, the learning algorithms aim to categorise between translated and non-translated class for each text. The performance of the learning model is evaluated in two modes: first, using the 10-fold cross-validation technique, and second, using a test dataset[8]. The results of the learning system are reported in terms of accuracy, precision, recall, and F-measure score[9], typical performance metrics.

[6]Details regarding the part-of-speech tagger and the parser employed in the present research are reported in Section 3.2.2.

[7]The format required is presented in Section 3.3.2.

[8]More details on evaluation techniques in Section 3.3.2.

[9]These evaluation metrics are explained in Section 5.2.1.2.

The results yield distinct types of knowledge representations according to the algorithm used, and are further presented in Chapter 5. Some classifiers also provide user-friendly knowledge representations, outlining patterns or decision trees, such as the JRip (Cohen, 1995) and J48 learning algorithms. The J48 classifier is a Weka implementation of the C4.5 algorithm outlined by Quinlan (1993), and is in general referred to as the Decision Tree classifier.

4.3.1.1 Classification of the Experiments

These stages are valid for any learning model discussed in the present thesis. As there are distinct learning models, comprising different data representations, a classification of experiments is outlined below. The experiments are organised into a set of *research scenarios*, and are numbered as follows:

1. A comparison between the learning model which uses all the translationese features available for that language except the simplification features under investigation and the learning model which uses all the indicators including the simplification features. In this thesis, this model is also referred to as the *'excluding simplification learning model'*. This comparison aims at identifying to what extent the simplification features do influence the learning model built with the remaining features in the data representation.

2. A comparison between the learning model which uses all the translationese features available for that language except the explicitation features under investigation and the learning model

which uses all the indicators including the explicitation features. In this thesis, this model is also referred to as the *'excluding explicitation learning model'*. This comparison aims at identifying to what extent the explicitation features do influence the learning model built with the remaining features in the data representation.

3. Simplification learning model: data representation comprises only the features proposed for simplification. These experiments aim at investigating whether the learning model is able to categorise the texts as translations or non-translations solely using the potential simplification features.

4. Explicitation learning model: data representation comprises only features proposed for explicitation. These experiments aim at investigating whether the learning model is able to handle the same task solely using the potential explicitation features.

5. Ablation study: data representation comprises only one feature at a time. It investigates to what extent each feature can distinguish between the two classes involved in the learning model.

The same set of experiments are reported for both languages, and their results are discussed in Chapter 5.

At this point, it is important to clarify the terminology used in this thesis when referring to distinct learning models. Because translationese refers to all the specific characteristics of translational language, whereas simplification and explicitation attempt to group these characteristics, the learning model which uses all the features available for that language is

hereafter referred to as *translationese generic learning model*[10]. Naturally, the generic learning model can include features which can be seen as also supporting a distinct translational hypothesis, such as simplification or explicitation. Thus, the simplification learning model and explicitation learning model can also be seen as translationese learning models.

To keep a consistency throughout the thesis, the model which uses all the features is the translationese generic learning model, whereas the other models are referred to according to the hypothesis investigated.

The first two research scenarios aim at identifying to what extent the simplification and explicitation features, respectively, influence the generic learning model. The *assumption* is as follows: if the addition of the simplification or explicitation features to a learning model leads to a higher performance of that learning model, then this can be interpreted as an argument for the existence of the corresponding hypothesis. Selecting a set of features F for one hypothesis H at a time is preferred, to inquire into the validity of that hypothesis H in terms of that set of features F.

To assess the statistical significance of the improvement brought to the machine learning system when including simplification or explicitation features over the learning system without these features, the paired two-tailed t-test is applied with a 0.05 significance level.

The next two scenarios aim at identifying to what extent the potential simplification- and explicitation features are able to distinguish between translations and non-translations without considering any other additional

[10]To ease understanding, the terminology is occasionally referred to as the *generic learning model*.

feature in their learning process. Furthermore, the ablation study analyses to what extent each feature is able to accomplish the same task on its own.

4.3.1.2 Learning Algorithms

The classifiers applied in the learning models are the following: JRip, Decision Tree, Naïve Bayes, IB1 and SVM (Witten et al., 2011). The evaluation results are outlined in the next chapter. These particular algorithms are chosen because the JRip and Decision Tree classifiers provide a knowledge representation easy to understand for humans, whereas Naïve Bayes has good results on text categorisation task, IB1 is known for its ability to handle well numerical attributes and SVM usually achieves high performance.

A meta-classifier is also employed, namely the Vote meta-classifier with the Majority Voting combination rule, which relies on the output of other classifiers. Unless specifically stated in the research scenario, the following three algorithms are used in the Majority Voting rule for both Spanish and Romanian experiments: SVM, IB1 and JRip classifiers output.

Since there are three distinct hypotheses, translationese, simplification and explicitation, the features used within the learning models differ. One aspect is further emphasised, the fact that the attributes used to support translationese, simplification or explicitation hypothesis are characteristics proposed within the literature or are inferred from the existing studies, and they are not absolute factors.

In the next subsection, the features used in the data representations are detailed.

4.3.2 Translationese Generic Learning Models

The most challenging part of the research on translationese, or the translational hypotheses, is to ascertain the potential features that are suspected to be specific to translational language. The lack of clear and precise explanations and hypotheses within the domain is the main issue which arises at this point. However, the corpus-based studies reported in the literature offer valuable clues for the selection of features for the learning model.

Besides the fact that machine learning techniques allow the investigation of more features at the same time, most of them are also able to handle noisy or irrelevant features for the sought task. For this reason, in the learning model reported in this thesis, there are several potential features for translationese, simplification and explicitation universals, included in the learning process.

The selection aims only at providing a solid set of attributes from which the learning model can extract the most reliable ones to classify the instances with a good performance.

In the selection of the features, the following types are given priority:

- attributes which are expected to be present in both types of text;
- attributes which are multilingual (i.e., attributes which can be computed for several languages, and not only available for one language);

- attributes which were previously suggested as being distinctive or relevant to one class or another, i.e., translated or non-translated class.

To ensure that the system learns the right target concept, and to prevent it from learning to classify according to the topic of a text, the current approach avoids the bag-of-words model[11]. Also, to ensure that the model is multilingual (i.e., it comprises only features which are available for more languages), the use of n-gram features is avoided as well.

Note, however, that the linguistic system slightly differs from one language to another, which results in a slightly distinct data representation for the Spanish and the Romanian learning frameworks. Nevertheless, the learning systems explored are largely similar. The parser and the part-of-speech tagger annotate the words, their lemma and their morphological information, and mark the sentences found in the corpora[12].

4.3.2.1 Remarks on the Hypotheses Investigated

Before enumerating the features used in the generic learning models for the two languages, a few remarks regarding translationese, simplification and explicitation are shortly outlined.

In this thesis, the notion of translationese is used in a neutral sense, referring only to the translation-specific language, without any

[11]The bag-of-words model, a well-known model in the natural language processing domain, is a model which represents a text as an unordered collection of words (Salton and McGill, 1983). Basically, for a given text, each feature of the data representation represents the frequency of occurrence of each word.

[12]Note that a word is seen as a sequence of characters in between adjacent blank spaces, and a sentence as a sequence of words found between two adjacent punctuation marks that end sentences. A token can be any type of punctuation mark or a word.

kind of negative implications. Second, the simplification and explicitation hypotheses are studied without considering their hypothesised subconscious tendency to explicitate or to simplify the message translated. These aspects require a distinct research method, and they are not the goal of this thesis.

A third aspect is further emphasised. According to several scholars, translation universals are seen as hypotheses regarding the nature of translational language, highlighting potential characteristics of this type of text. However, in this thesis, a slightly different view is preferred. Because translationese refers to all specific features of translational language, irrespective of type of feature, then in this thesis, translation universals are seen as sub-hypotheses which attempt to cluster distinct sub-types of these translationese characteristics. In other words, the set of translationese features is the union of all the features proposed in the study of translation universals in the literature and some other features, maybe undiscovered yet.

As a consequence, in this thesis, the features proposed to stand for simplification and explicitation are also studied separately, in distinct learning scenarios, thus creating the simplification and explicitation learning models. The features used in these specific learning models are outlined in Section 4.3.3 and 4.3.4, respectively.

The experiments and their specific features used are further reported in chronological order: the experiments on the Spanish corpora are initially presented, as these were conducted first, followed by the learning models for the Romanian data.

4.3.2.2 Data Representation for the Spanish Model

The translated features specific to translated text are chosen on the basis of the well-known assumption that translations exhibit their own specific lexico-grammatical and syntactic characteristics (Borin and Prütz, 2001; Hansen, 2003; Teich, 2003), "fingerprints" known as translationese. Therefore, analysing the morphological and syntactical features which can be extracted with part-of-speech taggers or parsers may lead to potential characteristics within which a translated text can be identified. In addition, a set of features of simplification and explicitation previously discussed in the literature are also included.

The translationese generic learning model for Spanish exploits twenty-two multilingual features in total. The potential translationese features, including the simplification and explicitation characteristics, are the following:

- proportion of nouns in a text;
- proportion of finite verbs in a text;
- the proportion of auxiliary verbs in a text;
- the proportion of adjectives in a text;
- the proportion of adverbs in a text;
- the proportion of numerals in a text;
- the proportion of pronouns in a text;
- the proportion of prepositions in a text;

- the proportion of determiners in a text;
- the proportion of conjunctions in a text;
- the proportion of grammatical words in a text;
- the proportion of grammatical words to lexical words[13];
- the average sentence length;
- the sentence depth;
- the proportion of simple sentences in a text;
- the proportion of complex sentence in a text;
- the proportion of zero sentences[14] in a text;
- the average number of senses per word;
- the average word length;
- the lexical richness of a text as the proportion of lemma types[15] to tokens;
- the information load of a text as the proportion of lexical words to tokens;
- the proportion of sentences which have at least one relative pronoun in a text[16].

[13]Some of these features are interrelated characteristics used in the model. For instance, the current feature and the proportion of grammatical words in a text. However, the performance of the learning model is not influenced by this interrelation.

[14]See details in Section 4.3.3.

[15]The number of unique lemmas in a text. This notion is explained in Section 4.3.3.1.

[16]I would like to express my gratitude to Georgiana Marşic for her useful suggestion on this particular feature.

Grammatical words, also known as function words, are represented by: determiners, prepositions, auxiliary verbs, pronouns, and conjunctions. Lexical words, also known as content words, are represented by nouns, verbs, adjectives, adverbs, and numerals.

Apart from the morphological classes listed above, features which are assumed to be relevant in the learning process, the specific features to stand for the simplification and explicitation hypotheses are further justified in Section 4.3.3 and 4.3.4, respectively. As previously pointed out in Chapter 2, some of these features have been analysed in corpus-based investigations on different translational hypotheses in Corpas, Mitkov, Afzal and Pekar (2008); Corpas, Mitkov, Afzal and García (2008) as well as in Corpas (2008).

4.3.2.3 Data Representation for the Romanian Model

Because the investigation of translationese on Romanian included a time-consuming task, namely to compile the necessary Romanian translational comparable corpora, the machine learning experiments on the Romanian data are conducted after the Spanish ones. This also had the advantage of having an idea of how the learning model performs using the feature vector reported above for the Spanish framework.

Building on the findings acquired from the Spanish experiments, the Romanian model employs a higher number of features, using also morphological sub-categories. Consequently, the translationese generic learning model designed for Romanian exploits thirty-nine multilingual features.

The extraction of the features is made using the part of speech tagger's output, with one exception. The only feature that involved a different type of processing is the proportion of the verbs which have an anaphoric zero pronoun in their subject position, an attribute which is further described in Section 4.3.4.3. A dependency parser for Romanian is not available; hence, the part-of-speech tagger presented in Chapter 3 is employed for these experiments.

Emphasising that translationese can be analysed at the morphological level of the texts (Laviosa, 2002; Toury, 1995), and also considering the features discussed in the literature for simplification and explicitation hypotheses, the forty-seven attributes of the translationese generic learning model for Romanian are listed below:

- the proportion of nouns in a text;
- the proportion of verbs in a text;
- the proportion of adjectives in a text;
- the proportion of adverbs in a text;
- the proportion of numerals in a text;
- the proportion of pronouns in a text;
- the proportion of adpositions in a text;
- the proportion of determiners in a text;
- the proportion of articles in a text;
- the proportion of conjunctions in a text;

- the proportion of grammatical words in a text;
- the proportion of grammatical words per lexical words in a text;
- the proportion of interjections in texts in a text;
- the proportion of proper nouns in texts in a text;
- the proportion of common nouns in texts in a text;
- the proportion of verbs in the first person plural in a text;
- the proportion of verbs in the first person singular in a text;
- the proportion of verbs in the second person plural in a text;
- the proportion of verbs in the second person singular in a text;
- the proportion of verbs in the third person plural in a text;
- the proportion of verbs in the third person singular in a text;
- the proportion of auxiliary verbs in a text;
- the proportion of modal verbs in a text;
- the proportion of verbs in the indicative mood in a text;
- the proportion of verbs in the subjunctive mood in a text;
- the proportion of verbs in the imperative mood in a text;
- the proportion of verbs in the infinitive mood in a text;
- the proportion of verbs in the gerund mood in a text;
- the proportion of verbs in the participle mood in a text;

- the proportion of comparative adjectives in a text;
- the proportion of positive adjectives in a text;
- the proportion of superlative adjectives in a text;
- the proportion of demonstrative pronouns and adjectives in a text[17];
- the proportion of indefinite pronouns in a text;
- the proportion of possessive pronouns in a text;
- the proportion of reflexive pronouns in a text;
- the proportion of negative pronouns in a text;
- the proportion of personal pronouns in a text;
- the proportion of interrogative negative pronouns in a text;
- the lexical richness of a text;
- the average sentence length;
- the average word length;
- the proportion of simple sentences in a text;
- the proportion of complex sentences in a text;
- the information load of a text;
- the proportion in a text of the verbs which have an anaphoric zero pronoun in their subject position[18];

[17]The tagger does not distinguish between them, so the feature considers both the demonstrative pronouns and the demonstrative adjectives.

[18]A definition and examples of this feature are further detailed in Section 4.3.4.2.

- the proportion in a text of sentences which have at least one relative interrogative pronoun[19].

One difference that is to be noted is that, for Romanian, grammatical words are represented by a slight distinct set of part of speech classes – determiners, articles, prepositions, auxiliary verbs, pronouns, conjunctions, and interjections – whereas the lexical word class is the same as for Spanish, being represented by nouns, verbs, adjectives, adverbs, and numerals.

The justification for the potential simplification and explicitation features is presented further in Section 4.3.3 and 4.3.4.

4.3.3 Simplification Learning Models

To assess to what extent the simplification features influence the learning model, these attributes are avoided in the main translationese model, being added at a later stage and their impact on the emerging learning model being analysed further.

The data representation for the Simplification Learning Model includes a set of previously proposed simplification features (Laviosa, 1998; Corpas, 2008). As there are two sets of experiments for each language, these features differ slightly due to availability of specific natural language processing tools.

[19]The part-of-speech tagger does not separate the relative pronoun from the interrogative one, marking both of them as relative interrogative pronouns.

4.3.3.1 Spanish Learning Model

For Spanish, the data representation for the Simplification Learning Model comprises the following features:

- average sentence length;
- average sentence depth, where the sentence depth is seen as the maximum depth of a syntactical tree;
- proportion of simple sentences in a text;
- proportion of complex sentences in a text;
- proportion of sentences without any finite verb in a text[20];
- the average number of senses per word[21];
- average word length, where the word length is computed as the proportion of syllables per word;
- lexical richness computed as the proportion of lemma types to number of tokens in a text;
- information load computed as the proportion of lexical words to tokens.

The justification for including these features in the model is provided in the following paragraphs.

[20]Throughout this thesis, this attribute is also referred to as the proportion of *zero sentences* in a text.

[21]Note that the ambiguity parameter is obtained by exploiting the Spanish Wordnet synsets (Verdejo, 1999).

Sentence Length

The *average sentence length* is largely investigated for both the simplification and the explicitation hypothesis, as it is hypothesised that short sentences would be interpreted in favour of simplification (Malmkjaer, 1997; Laviosa, 2002), whereas longer sentences present in translated texts would lead to an interpretation in favour of explicitation (Frankenberg-Garcia, 2004).

Apparently, regardless of what hypothesis is assigned, it appears to be distinctive and thus relevant in a classification task using machine learning techniques. Nevertheless, the rationale for including this attribute in the Simplification Model is as follows: short sentences tend to be easy to follow, and consequently more readable. In addition, splitting long and complex sentences is a known strategy in translation, and it is expected to result in a higher number of sentences, presumably also of shorter length. Thus, it is assumed that a lower value of the computed average will be obtained in translated texts. This will also lead to a potentially reliable feature in the categorisation task between translated and non-translated texts.

Sentence Depth

An alternative metric for the sentence length is the *sentence depth*, which is computed as the average value of the maximum depth of the syntactical trees given as an output from the dependency parser. This metric is previously proposed and analysed in a corpus-based approach by Corpas (2008); Corpas, Mitkov, Afzal and Pekar (2008).

It is undertaken in this machine learning approach to investigate whether it influences the learning model on the classification task. This attribute may prove to be distinctive because its value is expected to be lower in translations than in non-translated texts, an assumption based on the translators' known strategy of splitting long and complex sentences from the source text.

Sentence Type

The next three features are the proportions of *simple, complex and zero sentences* present in a text. A *zero sentence* is understood as a sentence which has no finite verb, seen as a sub-category of simple sentences. In this case, a *simple sentence* is the sentence with precisely one finite verb, whereas a *complex sentence* is the sentence with two or more finite verbs. The finite verb is indicated in the XML output of the Connexor parser (see the description of the parser in Chapter 3).

These three types of sentence are features dependent of each other, and the reason for considering all three of them is because this provides a clearer overview regarding their behaviour. This thesis points out that the opposed attribute should also be considered. According to the simplification hypothesis, simple sentences should appear predominantly in translated texts as opposed to non-translated texts, hence constituting a reliable attribute in the learning model.

Average Number of Senses per Word

The *average number of senses per word* attribute is aggregated in the learning model. Throughout this thesis, this attribute is also referred to

as the ambiguity feature (Corpas, 2008). Although it may appear as a controversial feature for the simplification hypothesis, note that most of the classifiers can discard irrelevant features in their training, and hence, its presence in the model should result in no harm.

Nevertheless, it is included in this set of simplification features following this rationale: on the one hand, the more numerous the senses of a word, the more ambiguous the message conveyed can appear; the more ambiguous the message, the more context is required for the reader to process and understand the message (Harley, 2008; Temnikova, 2012). On the other hand, translated texts are hypothesised to be easy to read and simple to understand.

It is thus assumed that translated texts, being simpler, would have a lower degree of ambiguity than non-translated texts. As a result, this feature could prove to be a potential reliable attribute in the classification task, which is the reason for employing it in the learning framework.

Word Length

The *word length*, where the word length is computed as the proportion of syllables per word[22], is expected to have a lower value for translated texts since the shorter the words used in a text, the easier-to-follow that text becomes (Harley, 2008). Consequently, this feature may appear to be indicative in the classification task.

[22]The syllables are automatically retrieved using a Ruby library to hyphenate words for Spanish. The documentation of the package, Text::Hyphen, can be found at: `http://rubydoc.info/gems/text-hyphen/1.4.1/frames`.

Lexical Richness

The lexical features in terms of lexical richness and information load are included in the data representation. Simplification is found to be validated at lexical level across several corpus-based studies, the results yielding a lower lexical density and information load in translated texts (Laviosa, 2002; Corpas, 2008; Xiao et al., 2010).

The standard way to compute lexical richness is to calculate the proportion of the number of types to the number of tokens. The reason for calculating lexical richness in a slightly distinctive manner, by the use of lemma types instead of word types[23], is the following: in the type/token ratio, the morphological variants of the same word are counted twice, although from the lexical point of view they represent the same thing and they are perceived as repetitions by the readers. Consequently, the metric for lexical richness adopts the use of type lemmas in order to better reflect the vocabulary variety implied in texts: i.e., the two distinct word types 'book' and 'books' would be counted only once because both word types represent the same thing from the lexical perspective (Corpas, Mitkov, Afzal and Pekar, 2008; Corpas, Mitkov, Afzal and García, 2008; Corpas, 2008).

In other words, by type lemmas is understood the number of distinct lemmas, in our case from a text since the parameter is computed at the text level, whereas type words are seen as the number of distinct words. Given the example above, 'book' and 'books', word lemmas are 1 because both

[23]A reminder of the notions used is as follows: the number of word types in a text is given by the number of unique words used in a text, whereas the number of word tokens is basically the length of the given text as number of words, without taking into consideration whether any of these words are repeated (Oakes, 2012, p. 133).

words have the same lemma, whereas the number of word types is 2 because there are two distinct words. Using the natural language processing tools, the corresponding lemma for each word is marked, so the value for this feature can be extracted.

Information Load

The information load attribute is a similar metric to the lexical density as outlined by Stubbs (1986, p. 33), who computes it as the ratio between the number of lexical words and the total number of words. In this work, the information load is computed as the ratio between lexical words and total number of tokens retrieved in the pre-processing phase. This indicator is referred to as the information load in the literature (Olohan, 2004, p. 100) and it is previously investigated in various corpus-based studies on simplification (Laviosa, 2002; Corpas, Mitkov, Afzal and Pekar, 2008; Corpas, 2008).

Translated texts are expected to exhibit a lower degree of information load, and thus represent a reliable feature for the classification task. Similarly, lexical richness is expected to manifest in the same way in translational language. A lower value for lexical richness would result in a higher level of repetitions in texts, a characteristic often observed in translations. Thus, these two metrics are expected to be relevant in the learning model, based on the assumption that translated texts have a poorer vocabulary than non-translated texts.

Next, the simplification features used in the Romanian model are outlined.

4.3.3.2 Romanian Learning Model

Given that a dependency parser is not yet available for the Romanian language, the simplification features are slightly different from the Spanish ones. Note that the justification for inclusion in the learning model of each attribute is outlined in the above paragraphs.

The potential simplification features for Romanian are:

- the lexical richness as the proportion of type lemmas per tokens;
- the average sentence length as the proportion of number of words per sentence;
- the average word length in terms of number of characters normalised by the number of tokens;
- the number of simple sentences[24] normalised by the total number of sentences in texts;
- the number of complex sentences normalised by the total number of sentences in texts;
- the information load as the proportion of lexical words to tokens.

Next, the explicitation learning model for both languages is presented.

[24]Given that the tagger does not provide any syntactic information, the following algorithm has been employed to compute this feature: sentences with one or zero personal verbs are considered to be 'simple sentences'.

4.3.4 Explicitation Learning Models

As the notion of explicitation and explicitness is interpreted in various ways across research studies, a clarification regarding the viewpoint adopted in this research is necessary. The present thesis reports on a computational approach aiming at distinguishing between translations and non-translations, taking into consideration a set of features previously proposed to stand in favour of explicitation. Whether indeed these features grasp the explicit character of translation or not is not the aim of the study. Neither is it to establish whether the explicitness of one text or another results from the translation process, or from any other reason. This is a product-orientated investigation, and the source texts are not considered in the experiments.

The rationale for preferring a product-orientated analysis for the explicitation hypothesis is the following: if there is such a phenomenon in translated texts, namely explicitation, which, as suggested in the literature, is supposed to leave traits in terms of a set of potential explicitation features, then an automatic system should be able to make the distinction between translated and non-translated texts by relying on the suggested features. This hypothesis emphasises a certain characteristic of translated texts as opposed to non-translated texts, which makes the comparable corpora, and consequently a product-orientated study, a suitable resource and methodology for explicitation.

The following paragraphs introduce the potential explicitation features which are to be included in the Spanish learning model:

4.3.4.1 Spanish Learning Model

Most of the explicitation features used in this research are proposed and analysed within the literature (Becher, 2009; Laviosa, 1995):

- the proportion of conjunctions in a text;
- the proportion of adverbs in a text;
- the proportion of sentences which have at least one relative pronoun in a text;
- the proportion of pronouns in a text.

The reason for the selection of each of these parameters is further outlined.

Conjunctions

Explicitation is often investigated in terms of connectives in several example- and corpus-based approaches. According to Pasch et al. (2003), conjunctions are a sub-type of connectives through which cohesion is attained in texts. Being investigated in several studies, the assumption is that a higher level of conjunctions in translational language would represent a validation of the explicitation hypothesis.

The lack of clause connectives, and in this case, a lower value of this proportion in texts, would lead to several implicit relations between clauses/sentences which the reader has to resolve. Translational language is expected to have a higher usage of conjunctions, therefore the attribute is assumed to be influential in the learning model.

Adverbs

The reason for including among explicitation features the proportion of adverbs in a text is: if indeed translational language tends to overuse sentential adverbials, a feature investigated in the explicitation hypothesis, then this increase would probably be reflected in the overall proportion of adverbs in texts. It is thus assumed to be a potentially reliable feature in the classification task. As the parser does not annotate sentential adverbs, the entire class of adverbs are considered in the learning model.

Also, adverbs are a morphological class found to be overused in translations by Borin and Prütz (2001), and this reinforces the assumption that they are a relevant indicator in the classification task.

Pronoun Features

A level of redundancy is pointed out in translations as translators, in their attempt to render the message explicitly, tend to overuse lexical repetitions or to repeat redundant grammatical items. As a consequence, a lower frequency of pronouns is assumed to characterise translated texts (Laviosa, 1995). For this reason, there are two features attempting to quantify features related to pronouns: first, the proportion of pronouns itself, and second, the proportion of sentences which have at least one relative pronoun in a text.

In addition, the disambiguation of pronouns in translational language (Olohan and Baker, 2000; Pápai, 2004) is another aspect which contributes to a lower number of pronouns. The repetition of names and noun phrases in order to explicitly illustrate the message conveyed results in a lower

proportion of pronouns in texts. Therefore, this explicitation-feature has the potential to be an influential feature in the learning model.

4.3.4.2 Romanian Learning Model

In addition to the above explicitation features discussed for the Spanish model, the optional ellipsis, a sub-type of grammatical cohesion, is adopted in the data representation for Romanian. Only the Romanian learning model uses this indicator in its data representation because this feature is language- and domain-dependent and no available framework was found to automatically retrieve the anaphoric zero pronouns for the Spanish medical and technical texts.

As the subject of a sentence has an important role in the correct understanding of the message employed, the type of ellipsis chosen is thus the ellipsis of the subjects, by employing as feature in the learning model the frequency of anaphoric zero pronouns in a text. To the best of the author's knowledge, the investigation of anaphoric zero pronouns in translated and non-translated texts has not been the subject of any research study to date.

The potential explicitation features studied for Romanian are:

- the proportion of indefinite pronouns in a text;
- the proportion of possessive pronouns in a text;
- the proportion of reflexive pronouns in a text;
- the proportion of negative pronouns in a text;
- the proportion of personal pronouns in a text;

- the proportion of interrogative negative pronouns in a text;
- the proportion of sentences which have at least one relative interrogative pronoun in a text;
- the proportion of conjunctions in a text;
- the proportion of adverbs in a text;
- the proportion of pronouns in a text;
- the proportion of verbs which have an anaphoric zero pronoun in their subject position in a text[25].

It can be noted that the number of features for Romanian is slightly greater mainly because the morphological sub-categories for pronouns are detailed, leading thus to a larger feature vector for the explicitation learning model. Nevertheless, the reason these features are included is the same as for pronouns, being justified earlier.

An observation on the decision to include the proportion of sentences which have at least one relative interrogative pronoun in a text needs to be pointed out. In order to adapt the current data representation to the Spanish one, this feature also considers interrogative-relative pronouns. As the part-of-speech tagger does not enable the extraction of only the relative pronouns from texts, this feature is including as well the interrogative pronouns because this is the closest feature which can be obtained given the constraints of the tools available for Romanian. Although this may

[25]A definition and examples of this feature are further detailed. Also note that, to facilitate understanding, this feature may be referred to as the AZP verbs attribute.

include a degree of noise, machine learning algorithms should be able to handle it because the noise involved is consistent throughout the dataset.

In addition, the feature vector for Romanian benefits from the possibility to include the proportion in a text of the verbs which have an anaphoric zero pronoun in their subject position. This notion, as well as the reason to include it, are explained in the next paragraphs.

It is assumed that translational language has the tendency to fill out the elliptical constructions (Øverås, 1998; Pápai, 2004), which leads to a potentially influential feature in the learning model. As all the features are captured automatically, a system is implemented to identify a sub-category of ellipsis for the Romanian language, namely the anaphoric zero pronouns, hereafter referred to as AZP .

The retrieval of the verbs which have zero pronouns in their subject position is done through a machine learning approach, and a detailed description of the system is given by Mihăilă, Ilisei and Inkpen (2011).

As the detection of the anaphoric zero pronouns is domain-dependent, the learning model used is carefully selected to be trained on the same journalistic domain with texts written in the same time-frame as the ones included in the Romanian corpus. The AZP identification model obtains an accuracy of 74%, being a sufficiently reliable system for the inclusion of the feature in the learning model.

A brief overview of the anaphoric zero pronouns in Romanian is outlined below.

4.3.4.3 Romanian Zero Pronominal Anaphora

A zero pronoun (ZP) is the gap in the sentence which refers to another entity in the text, which provides the necessary information for the correct understanding of the gap. The identification of the verbs which have zero pronouns in their subject position has been employed, and their proportions in texts constitute a feature integrated in the explicitation learning model for Romanian.

Although there are different classifications of ellipsis for Romanian (Mladin, 2005), the definition adopted in the AZP learning model is the following: an anaphoric zero pronoun appears when an anaphoric pronoun is omitted but nevertheless understood (Mitkov, 2002).

Also, there are two types of elliptic subject: zero subjects and implicit subjects (zero pronouns). The difference between them consists in the following aspect: zero pronouns can be lexically retrieved (ex. 1), while zero subjects cannot. If the zero pronouns would be present in texts, they would anaphorically refer to another entity.

In the following examples, quoted from Mihăilă, Ilisei and Inkpen (2011), these two types are illustrated: note that where the zero pronoun could be placed in the sentence is marked with $_{zp}[]$, whereas the zero subject is marked with the ⊘ sign (ex. 2).

1. $_{zp}[]$ A aterziat pe lună.
   [He/She] landed on the Moon.

2. ⊘ Afară plouă.
   [It] is raining outside.

The pronouns 'he/she' appear in the English translations because English does not allow the omission of the subject, whereas the correct understanding of the anaphoric zero pronoun is inferred from the context of the text.

Although the model employs the use of explicitation and simplification features, the main goal is to validate whether translated texts can be distinguished from non-translated texts, regardless of whether the features used in the learning model stand for one sub-hypothesis or another. Thus, in the last scenario, a learning model uses all the features available for the corresponding language.

Provided that all the features are presented, the following subsection outlines the assumptions of this framework.

4.4 Assumptions of the Learning Models

It is important to emphasise the assumptions considered when building these learning models.

First, it is assumed that there are such features which are distinctive for translational language (i.e., translationese exists), and are thus able to distinguish between translations and non-translations. A consequence of this assumption is that these features are then presumably universal: i.e., features which are able to detect translations in any circumstance, regardless of source or target language, regardless of genre, or any type of classification. Yet, this thesis only considers the first part of the assumption, and not its consequence: i.e., it is presumed that there may be

a set of features which can distinguish translated and non-translated texts in a given context, or for a given language. This thesis does not investigate the universal aspect of the features under examination.

Second, if a set of features is to classify texts in the translated and non-translated categories, then those features are considered to be characteristic of translationese, either favouring one class or another (i.e., translated or non-translated texts). For instance, if the feature 'word length' is hypothetically found to be a reliable feature in the learning model, then it is considered that 'word length' may be a feature of translationese, subject to being corroborated in similar research studies with the same methodology in order to be then seen as a 'universal feature' of translational language.

The third assumption is that the target concept of the learning model (i.e., translationese) can be represented by a conjunction/disjunction of features. By considering a set of translationese, simplification and explicitation features in the learning process due to suggestions pointed out in the literature, the model is biased in the following sense: considering these or a subset of these features, the learning process can learn to identify translated from non-translated texts.

4.5 Conclusions

Considering an under-explored approach within the domain of translation studies, this chapter starts by motivating the turn taken in this research when choosing the machine learning approach, which offers a lot of potential in discovering new knowledge taking the form of patterns retrieved in the

learning process. Further strengths of this approach are then outlined in the context of translational language research, and the preliminary notions regarding such a framework are introduced.

Section 4.3 reports on the fundamental elements of the multilingual translationese learning model proposed in this thesis. Data representation justifies the inclusion of each feature within the learning model. The current research combines these features together so that the influence of a certain group is analysed. The features are organised according to the proposed simplification and explicitation features present in the literature, designing two other learning models called the simplification and explicitation learning model.

Finally, the section emphasises and discusses the assumptions considered in the learning framework. In the next chapter, the findings retrieved from the learning models are reported.

Chapter 5

Evaluation

5.1 Overview

The objective of this research is to analyse translationese and two of its related sub-hypotheses by implementing a system capable of learning to distinguish between translated and non-translated texts. The experiments use the Spanish and Romanian resources described in Chapter 3, and consequently, the evaluation of the results is divided into two parts: first the Spanish experiments, and then the Romanian experiments.

The simplification and explicitation sub-hypotheses are also analysed as follows: first, by comparing the translationese generic learning model with the model which excludes the simplification or explicitation features from its data representation. Second, by creating dedicated learning models for each sub-hypothesis, namely *simplification learning model* and *explicitation learning model*, in order to assess to what extent the learning

model is able to classify the texts relying only on the corresponding potential features.

Section 5.2 reports on the Spanish experiments starting with the translationese generic learning model. Then the comparison between the generic model and the learning model which excludes simplification from its data representation is reported in Section 5.2.1. This comparison scenario analyses to what extent the simplification features influence the performance acquired by the translationese generic learning model. The same type of comparison and corresponding analysis are conducted for the explicitation sub-hypothesis in the same section. Detailed results are reported for both learners and the section ends with an additional experiment: the same comparison for both sub-hypotheses analysed on medical and technical domains. This experiment is presented in Section 5.2.2.3, and it aims to investigate to what extent the classifiers are learning to categorise on each of these domains separately.

The simplification learning model is reported in Section 5.2.3, whereas the explicitation learning model is discussed in Section 5.2.4. These two learning models are subjected to an additional experiment to evaluate to what extent the target concept is learnt on the two domains considered: medical and technical.

The last component of the section on the Spanish experiments is the ablation study, reported in Section 5.2.5.

The Romanian experiments are described in Section 5.3, keeping the same structure as for the Spanish ones. The translationese generic learning model is presented in Section 5.3.1. The comparison between the learning

models is discussed in Section 5.3.2. Then, the simplification learning model, the explicitation learning model, as well as the ablation study are reported in the next three sections of the chapter.

Section 5.4 provides a discussion of the outcomes obtained for both the Spanish and the Romanian data, and it includes a comparison to related work in the field. The chapter findings are summarised in Section 5.5.

5.2 Spanish Experiments

The comparable corpus for Spanish employed for these experiments is described in Chapter 3. The experiments are trained on the entire dataset, regardless of domain, and evaluated in two ways: first, using the 10-fold cross-validation technique highly used in machine learning, and second, using a chosen randomly generated dataset from all three pairs of comparable texts.

The training dataset comprises 450 instances, with 156 instances for the translated text class and 294 for the non-translated text class. The randomly chosen test dataset comprises 148 instances: 52 for the translated class and 96 for the other. Note that the number of instances in both datasets is balanced among the type of texts: i.e., medical texts written by professionals, medical texts written by students, and technical texts. In this way, the learning process avoids the tendency to learn the target concept for a certain type of text. The testing dataset has thus the following number of instances, according to each pair of the corpora:

- Set pair one: MTP vs. MTPC (2 translations vs. 2 non-translations);

- Set pair two: MTS vs. MTSC (36 translations vs. 66 non-translations);
- Set pair three: TT vs. TTC (14 translations vs. 28 non-translations).

Additionally, the learning model is also evaluated on separate datasets corresponding to each corpus domain: medical and technical, respectively. As the first pair has only 4 instances, representing a very small amount of instances, the results obtained would not be relevant for the experiment. Thus, only the latter two sets are investigated in order to analyse the performance obtained by the learning model on distinct domains.

The learning classifiers used for each learning scenario are the following: a baseline represented by the ZeroR classifier, a rules algorithm represented by JRip (Cohen, 1995), a decision tree algorithm represented by J48 (Quinlan, 1993), a bayes classifier represented by the well-known Naïve Bayes (John and Langley, 1995), a function classifier represented by the SVM learner (Platt, 1999), a lazy learning algorithm represented by the nearest-neighbour classifier also known as IB1 (Aha and Kibler, 1991), and the Majority Vote meta-classifier (Kuncheva, 2004). Details regarding these learning algorithms are provided below.

Unless otherwise specified, the meta-classifier takes the decision for each instance tested according to the majority vote among a set of three other classifiers: SVM, IB1 and JRip algorithms. These three classifiers have been selected because of two aspects: first, because it is important to aggregate distinct types of learning algorithms in the voting process, and in this case the categories involved are: the lazy learning classifiers, the rule classifiers, and the function classifiers. The second reason is the

observed high accuracies obtained by this particular algorithm within these categories, which thus suggest it would achieve better results in the voting process.

One observation needs to be made at this point: most of the classifiers do not use their default settings in Weka. These are adjusted in order to obtain better results, based on the cross-validation evaluation and on the training data, while the test data remains unchanged (since it is reserved only for the final testing). All the adjusted arguments of the classifiers are as follows[1]:

- SVM, which implements the sequential minimal optimization algorithm for training a support vector classifier (Platt, 1999), uses logistic models for probability estimates;

- J48, the decision tree classifier employed (Quinlan, 1993), has the confidence factor used for pruning of 0.01 to incur more pruning to the output;

- JRip, the classifier that implements a propositional rule learner (Cohen, 1995), has the minimum total weight of the instance in a rule of 10;

- Naïve Bayes, the algorithm which uses estimator classes in its learning, employs a kernel estimator for numeric attributes rather than normal distribution (John and Langley, 1995);

[1]The other settings, not mentioned in this list, are the default ones that Weka uses.

- Vote meta-classifier combines the SVM, the IB1, and the JRip classifiers with the corresponding settings mentioned above (Kuncheva, 2004).

All the attributes employed in the learning models are numeric, to facilitate the use of the classifiers chosen. The learning system for the Spanish data exploits twenty-two multilingual features in total, as presented in Chapter 4. Given that some of these attributes are also considered to be potential attributes for specific translational sub-hypotheses, not only for translationese, the investigation of simplification and explicitation sub-hypotheses is also conducted.

Unless otherwise specified, throughout all the experiments the performance of the classifiers used is reported for two types of evaluation modes: the 10-fold cross-evaluation, and the test dataset evaluation.

The next sections report the results obtained for various learning models, having the purpose to investigate translationese, simplification and explicitation hypotheses. Note that specific background information regarding the notions or metrics used is given as needed.

5.2.1 Translationese Generic Learning Model

To assess whether the chosen classifiers would perform better if the features proposed within the model were filtered first, as a pre-processing stage, the algorithms evaluate the model in both cases: with all the attributes, and with the remaining attributes after considering Chi-squared filter ranking.

Spanish		
Classifier	*10-fold cross-validation*	*Test set*
Baseline	65.33%	64.86%
Naïve Bayes	76.00%	83.11%
JRip	78.22%	80.41%
J48	79.78%	80.41%
IB1	76.44%	82.43%
SVM	81.11%	83.78%
Majority Vote	**85.11**%	**87.16**%

Table 5.1: Generic Learning Model: Classification Accuracies.

In Table 5.1, the results of the classifiers are provided for the learning model which considers all the attributes, the translationese generic learning model. On cross-validation, the learning model can distinguish between translated and non-translated texts with accuracies between 76% and 85.11%, values above chance level. On test evaluation, the results appear to be slightly higher, reaching up to 87.16%. The baseline classifier, ZeroR, considers the majority class from the dataset. Therefore, the value for the baseline is 64% - 65% accuracy because the majority class is non-translated. The balance of 2:1 kept between translated and non-translated class at the compilation stage of the resource also sets the baseline.

Next, the ranking of the features provided by two algorithms is reported.

5.2.1.1 Feature Ranking

Seeking a specific ranking of the attributes, the output of the feature selection evaluators is further analysed. The Information Gain and Chi-square algorithms provide the ranking reported in Figure 5.2, from the

highest rank to the lowest. The table excludes the null-valued attributes, marking these in italic in the Appendix in Table B.2.

Spanish Data	
Information Gain	**Chi-squared**
lexicalRichness	lexicalRichness
finiteVerbs	finiteVerbs
numerals	numerals
adjectives	adjectives
sentenceLength	sentenceLength
prons	prons
simpleSentences	wordLength
wordLength	simpleSentences
grammaticalWords	zeroSentences
zeroSentences	nouns
nouns	infoLoad
infoLoad	grammaticalWords
...	...

Table 5.2: Attribute Ranking Filters for the Generic Learning Model.

In general, it can be noted that the two feature selection algorithms acquire approximately the same knowledge, particularly for the top six attributes. As the slight variation is minimal, the most interesting part is that lexical richness, one of the well discussed features of simplification, appears right at the top of the ranking list, indicating that it is a highly relevant feature in the learning process. The fifth, seventh and eighth places are taken by the following potential features of simplification: sentence length, simple sentences and word length.

After the removal of the attributes which were scored as having 0 values after their evaluation on the full training dataset, the learning model employs the same classifiers using only the attributes mentioned in Table 5.2. In Appendix B.1, the accuracies of the new learning model that excludes the attributes with zero scores are detailed. The results obtained

on 10-fold cross-validation are similar, but most of the algorithms report slightly lower values. This indicates that the attributes that are removed bring a small contribution in the learning process. The lowest value is obtained by Naïve Bayes, having an accuracy of 75.56%, and the highest by the Vote meta-classifier, 80.89%. Note that the highest accuracy obtained, 80.89%, is considerably lower than the earlier corresponding performance obtained by the same meta-learner, which was 85.11%.

Since the classifiers appear not to need this pre-processing stage, hereafter the generic learning model selected for the experiments comprises all the features listed in Chapter 4.

5.2.1.2 Precision, Recall and F-measure Values

Before reporting the detailed results for the learning models, the necessary metrics are described at this point. These types of metrics are reported for all the learning models investigated in this thesis.

Precision is the number of correct results divided by the number of all returned results, whereas *recall* is the number of correct results divided by the number of results that should have been returned. In the current context, precision and recall for each class are outlined as follows: an algorithm's precision for a certain class A is the proportion of the identified instances that truly have class A to the number of instances that the algorithm classified as class A.

Recall indicates how much of class A is actually correctly predicted by the algorithm. It is the proportion between the number of instances which are labelled as belonging to class A and the total number of instances that

truly belong to class A (and should have been retrieved from the dataset). *F-measure*, traditionally the F_1 score, is the harmonic mean of precision and recall. All three metrics presented have values ranging between 0, as their lowest score, and 1, as their best score. They are formally represented as follows[2]:

- Precision=$\frac{TP}{TP+FP}$
- Recall=$\frac{TP}{TP+FN}$
- F-measure: $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

The detailed evaluation by class for the generic learning model for each classifier is reported in Table 5.3. The information is reported in terms of precision, recall and f-measure score for the 10-fold cross-validation evaluation. These scores emphasise which classifier better recognises a certain class.

According to Table 5.3, the overall results indicate that the non-translated class appears to be easier to predict than the translated class. The Voting algorithm achieves the highest f-measure scores for both classes, with 0.892 and 0.762 for the non-translated and the translated class, respectively.

In terms of precision and recall, the highest results are obtained by the same algorithm, the Vote meta-classifier. The detailed results for the translationese generic learning model indicate that the non-translated class

[2]The false positive, FP, is the number of instances incorrectly classified as belonging to class A. The true positive, TP, is the number of instances correctly classified as pertaining to class A. In contrast, the false negative, FN, is the number of instances incorrectly classified as not pertaining to class A.

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.653	1	0.79	non-translated
Naïve Bayes	0.688	0.564	0.62	translated
	0.789	0.864	0.825	non-translated
JRip	0.75	0.558	0.64	translated
	0.793	0.901	0.844	non-translated
J48	0.81	0.545	0.651	translated
	0.794	0.932	0.858	non-translated
IB1	0.658	0.667	0.662	translated
	0.822	0.816	0.819	non-translated
SVM	0.789	0.622	0.695	translated
	0.82	0.912	0.863	non-translated
Vote	**0.856**	**0.686**	**0.762**	translated
	0.849	**0.939**	**0.892**	non-translated

Table 5.3: Generic Learning Model. Evaluation mode: 10-fold cross-validation.

is identified better than the other one. There are, however, slight variations which suggest that the Vote algorithm has a slight edge in terms of precision for the translated class.

5.2.1.3 Translational Patterns

Knowledge representation of the patterns retrieved in the learning process can be illustrated in terms of a pruned decision tree, using a J48 classifier, and a rule set, using the JRip classifier.

Figure 5.1 shows the decision tree created by the J48 classifier for the translationese generic learning model. The goal of the decision tree is to illustrate how the features are used in the learning model in order to classify whether an instance belongs to the translated or non-translated class. The leaf nodes show which class is assigned, and the number in the brackets

indicates how many instances are assigned to that node. If there are two numbers in the brackets, then the second number indicates how many of those instances are incorrectly classified as a result.

```
lexicalRichness <= 0.16
|   sentenceLength <= 16.81: non_translated (30.0)
|   sentenceLength > 16.81
|   |   prons <= 0.05: translated (90.0/12.0)**
|   |   prons > 0.05
|   |   |   numerals <= 0.03: non_translated (15.0/1.0)
|   |   |   numerals > 0.03
|   |   |   |   finiteVerbs <= 0.09
|   |   |   |   |   zeroSentences <= 0.29: translated (9.0)
|   |   |   |   |   zeroSentences > 0.29: non_translated (3.0/1.0)
|   |   |   |   finiteVerbs > 0.09: non_translated (2.0)
lexicalRichness > 0.16: non_translated (301.0/67.0)**
```

Figure 5.1: Generic Learning Model: Pruned tree output from the Decision Tree classifier. Evaluation mode: 10-fold cross-validation.

The pruned tree output shows that the most informative feature is lexical richness, followed by the sentence length attribute and the proportion of pronouns in a text. Also, the numerals and the finite verbs, being on the next levels of the tree, make a valuable contribution in the classification.

For the translated class, most of the instances are classified using the first three levels of the decision tree, thus considering lexical richness, sentence length and pronoun attributes[3]. For the non-translated class, the leaf node mostly used is on the first level of the decision tree considering only the lexical richness.

The other classifier which provides an intuitive knowledge representation from the learning process is JRip. Before discussing

[3]Throughout the figures which illustrate translational patterns the double asterisk indicates the leaf node or the JRip rule that is most used for each of the classes.

the rules obtained by the classifier for each learning model, the specific output provided by this classifier is introduced.

Its pruned output indicates how the decisions are made, retrieving a rule set from the learning experience. Note that the number in brackets represents basically the same information as for the Decision Tree. For instance, the notation of one rule *(feature=1) => class=translated (10.0/3.0)* covers instances having total weights of 10.0, out of which instances with weights of 3.0 are misclassified. As the weight throughout all the experiments is 1, representing one instance, the rule given in the example is used for 10 instances, out of which 3 examples are incorrectly classified by the learner.

The rule set produced by JRip is an *if-else-then* set. The first rule which classifies an instance is also the only rule to classify that particular instance. There are not two rules to choose from in order to classify an instance.

Continuing to analyse the generic learning model through the lens of the JRip classifier, the rule set acquired is presented in Figure 5.2.

There are six rules obtained by this classifier. The first five rules identify the translated texts. If none of these rules apply for a given instance, then the last rule classifies it as non-translation.

Most translated instances get classified using the first rule, the one that uses lexical richness, sentence length, auxiliary verbs and word length in its decision. More precisely, 42 instances are correctly identified with this rule. The second rule considers the numerals, nouns, and sentence depth. It can be observed that the rules tend to aggregate distinct features in their

```
Rule 1: (lexicalRichness <= 0.16) and (sentenceLength >= 20.33)
and (auxVerbs <= 0.007422) and (wordLength <= 2.45)
=> class=translated (43.0/1.0)**

Rule 2: (numerals >= 0.05) and (nouns >= 0.34)
and (sentenceDepth >= 1.89) => class=translated (26.0/6.0)

Rule 3: (finiteVerbs <= 0.09) and (lexicalRichness <= 0.17)
and (conjs <= 0.04) => class=translated (35.0/9.0)

Rule 4: (sentAtLeastOneIntRelPron <= 0.191601) and
(sentenceDepth >= 1.88) and (auxVerbs >= 0.009782)
and (complexSentences >= 0.33) => class=translated (14.0/2.0)

Rule 5: (finiteVerbs <= 0.09) and (grammsWPerLexicsWords <= 0.603905)
and (dets >= 0.1) and (adjectives <= 0.1) and (adverbs <= 0.03)
=> class=translated (18.0/3.0)

Rule 6: => class=non_translated (314.0/41.0)**
```

Figure 5.2: Translationese Generic Learning Model: JRip classifier rules output. Evaluation mode: 10-fold cross-validation.

classification process, considering in total 11 of them. For the translated class, the first and the third rule identify most of the translated texts.

The next section illustrates the detailed results of the learning model when certain features are removed from the data representation.

5.2.2 Comparison between Learning Models

The machine learning system that includes the simplification features and the learning system whose data representation excludes these features are compared in order to assess the difference that occurs between them. The same comparison is also made for the explicitation features.

In order to assess the statistical significance, the paired two-tailed t-test has been applied with a 0.5 significance level for all the classifiers

employed. The statistically significant improvement is marked throughout these experiments with an asterisk by the accuracy value of the classifier.

The rationale for comparing these two learning models is based on the following *assumption*: *if the removal of the simplification or the explicitation features in the model leads to a lower performance of the learning process, then an argument is brought in favour of the corresponding hypothesis.*

Spanish Data						
Classifier	**Generic Data Representation**		**Excluding Simplification Model**		**Excluding Explicitation Model**	
	10-fold cv.	*Test set*	*10-fold cv.*	*Test set*	*10-fold cv.*	*Test set*
Baseline	65.33%	64.86%	65.33%	64.86%	65.33%	64.86%
Naïve Bayes	76.00%	83.11%	72.22%	79.73%	73.56%	80.41%
JRip	78.22%	80.41%	66.89%	74.32%	74.44%	79.05%
J48	79.78%	80.41%	70.44%	75.68%	79.11%	79.73%
IB1	76.44%	82.43%	70.89%	77.70%	77.11%	82.43%
SVM	$\diamond$ $\star$81.11%	83.78%	72.22%	77.70%	77.11%	81.76%
Vote	$\diamond$ $\star$**85.11**%	**87.16**%	**77.78**%	**83.11**%	**81.11**%	**85.81**%

Table 5.4: Comparison between the learning models: Accuracies for several classifiers.

Since there are no classifiers which registered a statistically worse performance compared to the learning model that excludes the simplification features, only the improvement is signalled: the $\diamond$ sign marks the statistical improvement when the values are compared to the excluding explicitation learning model, and the $\star$ sign is for the case when the accuracies are compared to the excluding simplification learning model.

The first observation on all the results obtained is that the Vote algorithm performs consistently better than all the other classifiers. Overall, the highest results are still the ones obtained by the generic learning model, having 85.11% on cross-validation evaluation and slightly better on test dataset, reaching 87.16% accuracy.

Slightly lower values are obtained for the same classifier on the excluding explicitation learning model. Also, according to T-test, the generic model obtains a statistically significant improvement compared to the latter model. This indicates that, although the values are slightly lower for this classifier, the difference is still statistically significant.

Compared to the excluding simplification learning model, the Vote meta-classifier for the generic learning model appears significantly better according to t-test evaluation. The results when simplification features are removed are clearly lower, registering 77.78% accuracy on 10-fold cross-validation and 83.11% on test evaluation.

Another algorithm which attains statistical significance for the generic learning model is the SVM classifier. It obtains statistically better accuracy when compared to both excluding simplification and explicitation learning models.

Although the results decrease for the other models, it has to be noted nevertheless that most of the classifiers are able to distinguish between translated and non-translated texts with accuracies above chance level.

At this point, another observation needs to be emphasised. With only one exception, all the classifiers exhibit lower accuracies for the other two learning models when compared to the generic model. The exception is the IB1 classifier for the excluding explicitation learning model when it registers a 77.11% accuracy, a value which is slightly greater than the one obtained in the generic model, namely 76.44%. This is an unexpected behaviour, and it indicates that the removal of the features helped the classifier achieve a slightly better result.

In comparison to the generic learning model, the exclusion of the simplification features leads to a decreased performance of the classifiers. The results range between 66.89% and 77.78% for the 10-fold cross-validation, and from 74.32% to 87.16% for the test evaluation. The biggest difference is noted for the JRip classifier, shortly followed by J48 and SVM. Moreover, the difference appears to be statistically significant for two of the learners employed, namely for the SVM and Vote learning algorithms.

Comparing the model which excludes explicitation features with the generic model, the classifiers tend to have smaller, yet similar results: the highest difference registered, more precisely of 4%, is for the SVM and Voting algorithms, which also signals that the difference is statistically significant.

To assess to what extent the classifiers perform for the detection of each class, more detailed information about the results is required. To this end, precision, recall and f-measure for each classifier are further explained for the two learning models: the model excluding simplification features, and the model excluding explicitation features.

5.2.2.1 Excluding Simplification Learning Model

Overall, the excluding simplification learning model obtains accuracies between 66.89%, for JRip, and 77.78%, for the Vote meta-classifier, on the 10-fold cross-validation evaluation. On the test evaluation, the accuracies range between 74.32% and 83.11%.

Since JRip has such a low value, dropping from 78.22% (accuracy acquired for the generic model) the result indicates that the learner was

relying on the simplification features. This tendency can also be noted from the ranking of the attributes shown in Figure 5.2 and, thus, the low performance is expected.

Precision, Recall and F-measure Values

In Table 5.5, the precision, recall and f-measure scores are reported for the excluding simplification learning model.

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.653	1	0.79	non-translated
Naïve Bayes	0.645	0.442	0.525	translated
	0.746	0.871	0.804	non-translated
JRip	0.529	0.41	0.462	translated
	0.72	0.806	0.761	non-translated
J48	0.593	0.468	0.523	translated
	0.746	0.83	0.786	non-translated
IB1	0.577	**0.603**	0.589	translated
	0.784	0.765	0.775	non-translated
SVM	0.648	0.436	0.521	translated
	0.745	0.874	0.804	non-translated
Vote	**0.775**	0.506	**0.612**	translated
	0.779	**0.922**	**0.844**	non-translated

Table 5.5: Excluding Simplification Learning Model. Evaluation mode: 10-fold cross-validation.

Given that the Vote classifier indicated the highest results for this learning model, it is expected, to some extent, to notice similarly high results for the partial results by class. Indeed, the Vote learner achieves the highest scores for almost all the metrics, with one exception: in terms of recall, for the translated class, the Vote algorithm is overtaken by the IB1 classifier, the latter obtaining 0.603.

The most impressive result is for the recall of the non-translated class obtained by the Vote meta-classifier, reaching a value of 0.922. Overall, for the translated class, the highest f-measure score is 0.612, whereas for the other class a value of 0.844 is attained. Note also the same tendency as for the generic learning model, i.e., having better results for the non-translated class.

Next, the translational patterns registered for this learning model are pointed out.

Translational Patterns

As the JRip classifier obtains an accuracy of only 66.89% on 10-fold cross-validation, the rules reported by the classifier are unreliable even though the learner obtained a better accuracy on the test evaluation. The 74.32% accuracy for that type of evaluation can be considered to appear by chance, given the low results on cross-validation.

```
Rule 1: (finiteVerbs <= 0.09) and (adjectives <= 0.09)
and (auxVerbs <= 0.005999) => class=translated (94.0/27.0)
Rule 2:  => class=non_translated (356.0/89.0)
```

Figure 5.3: Excluding Simplification Learning Model: JRip classifier rules output. Evaluation mode: 10-fold cross-validation.

Figure 5.3 reports its output. It can be observed that the decision is made on the basis of finite verbs, adjectives and auxiliary verbs in the rule employed to detect translated texts. Everything else is classified as non-translated, using the last rule.

```
finiteVerbs <= 0.09
|   adjectives <= 0.08
|   |   grammaticalWords <= 0.41
|   |   |   numerals <= 0.01: non_translated (4.0)
|   |   |   numerals > 0.01: translated (117.0/34.0)**
|   |   grammaticalWords > 0.41
|   |   |   dets <= 0.14: non_translated (10.0)
|   |   |   dets > 0.14: translated (3.0/1.0)
|   adjectives > 0.08: non_translated (202.0/59.0)**
finiteVerbs > 0.09: non_translated (114.0/12.0)
```

Figure 5.4: Excluding Simplification Features Model: J48 classifier pruned decision tree output. Evaluation mode: 10-fold cross-validation.

Considering the fact that the simplification features are removed in this learning model, the most reliable features, given the data representation, are reorganised. It can be observed that the only common attributes are the finite verbs and the numerals. The finite verbs appear on the first level of the decision tree, as the root node, being followed by the adjective attribute on the second level.

Analysing the decision tree for the translated class, the leaf node which has the highest number of correct classifications for the translated class, more precisely 83, considers finite verbs, adjectives, grammatical words and numerals in the classification. For the other class, the leaf node which has the highest number of correct classifications, namely 143, considers only the finite verbs and adjectives attribute, the first two levels of the tree.

The accuracy obtained using this decision tree is 70.44% for the 10-fold cross-validation and 75.68% for the test evaluation mode.

In the next section, the detailed results for the excluding explicitation learning model are presented.

5.2.2.2 Excluding Explicitation Learning Model

The overall learning model achieves accuracies between 73.56%, attained by Naïve Bayes, and 81.11%, obtained by the Vote classifier, on 10-fold cross-validation evaluation. On the other method of evaluation, on the test data, the model achieves accuracies between 79.73% and 85.81%.

The results presented show that the task of categorisation between the two types of text is achieved even without the features considered to stand for explicitation. The detailed results by class are discussed in the next paragraphs.

Precision, Recall and F-measure Values

In Table 5.6 the precision, recall and f-measure scores are reported for the excluding explicitation learning model.

In terms of f-measure scores, the highest values for both classes are achieved by the Vote meta-classifier. In terms of precision and recall, the best results are achieved by other classifiers.

The highest precision for the translated class is attained by the decision tree classifier, J48, obtaining a value of 0.810, whereas for the non-translated class, the IB1 learner obtained the top value, 0.819. Note that both values are outstanding.

In contrast, the values of recall tend to be lower for the translated class: IB1 obtains the best value compared to other learners, and it acquires a value of 0.654. For the non translated class though, the results are much

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.653	1	0.790	non-translated
Naïve Bayes	0.664	0.481	0.558	translated
	0.76	0.871	0.811	non-translated
JRip	0.688	0.481	0.566	translated
	0.762	0.884	0.819	non-translated
J48	**0.810**	0.519	0.633	translated
	0.786	**0.935**	0.854	non-translated
IB1	0.675	**0.654**	0.664	translated
	0.819	0.833	0.826	non-translated
SVM	0.709	0.577	0.636	translated
	0.796	0.874	0.833	non-translated
Vote	0.803	0.603	**0.689**	translated
	0.814	0.922	**0.864**	non-translated

Table 5.6: Excluding Explicitation Learning Model. Evaluation mode: 10-fold cross-validation.

higher overall and the decision tree obtains 0.935 recall, representing the highest value in this category.

The next paragraphs illustrate the patterns obtained in the learning process by the JRip and J48 classifiers.

Translational Patterns

In Figure 5.5, the JRip algorithm reports the rule set obtained in the learning process for this model. The classifier obtained an accuracy of 74.44% on 10-fold cross-validation and 79.05% on test evaluation using the three rules illustrated.

In the first rule, where lexical richness, finite verbs and word length are the features used, the learner correctly identified the translated text

```
Rule 1: (lexicalRichness <= 0.17) and (finiteVerbs <= 0.09)
and (wordLength <= 2.54) => class=translated (88.0/12.0)
Rule 2: (numerals >= 0.05) and (ambiguity <= 2.07) and
(preps <= 0.14) => class=translated (15.0/3.0)
Rule 3: => class=non_translated (347.0/68.0)
```

Figure 5.5: Excluding Explicitation Learning Model: JRip Rule Set. Evaluation mode: 10-fold cross-validation.

```
lexicalRichness <= 0.16
|   sentenceLength <= 16.81: non_translated (30.0)
|   sentenceLength > 16.81
|   |   numerals <= 0.03
|   |   |   dets <= 0.13
|   |   |   |   simpleSentences <= 0.37
|   |   |   |   |   lexicalRichness <= 0.06: translated (6.0/1.0)
|   |   |   |   |   lexicalRichness > 0.06: non_translated (23.0/4.0)
|   |   |   |   simpleSentences > 0.37: translated (5.0)
|   |   |   dets > 0.13: translated (8.0)
|   |   numerals > 0.03
|   |   |   grammaticalWords <= 0.41: translated (73.0/7.0)**
|   |   |   grammaticalWords > 0.41: non_translated (4.0/1.0)
lexicalRichness > 0.16: non_translated (301.0/67.0)**
```

Figure 5.6: Excluding Explicitation Learning Model: J48 classifier pruned decision tree output. Evaluation mode: 10-fold cross-validation.

for 76 instances of translated class. The second rule is used to a much lesser extent, detecting only 12 translated texts, by employing numerals, ambiguity and prepositions in the decision stage. The last rule, as expected, detects the non-translated class, correctly detecting 279 instances.

The pruned decision tree is illustrated in Figure 5.6. The classifier achieves an accuracy of 79.11% using this decision tree on the 10-fold cross-validation. The learner considers lexical richness as the root node of the tree, selecting sentence length as the second level of the tree. Numerals are shown on the third level, whilst determiners and grammatical words are placed on the fourth level.

For the translated class, the leaf node which classifies most of the instances makes use of grammatical words, numerals, sentence length and lexical richness. In contrast, for the other class, the learner decides for most of the instances using only lexical richness value.

An additional experiment was conducted. Since the corpora used for Spanish comprises two domains, the same comparison between the three learning models is analysed.

5.2.2.3 Evaluation on Medical and Technical Datasets

Given that the Spanish data is divided into two domains, an additional experiment evaluates the learning model on separate types of corpus: medical and technical domains. The sets of the comparable pairs of the resource were presented in Chapter 3.

The learning model is trained on the entire training dataset, but the classifiers are evaluated on two test datasets: one for the medical domain, and one for the technical domain. Given that the training dataset is balanced, keeping a proportion of 2:1 between translated and non-translated classes, the test dataset for each domain needs to maintain the same balance. As pair 1 has only 5 instances for both classes, this pair of comparable sub-corpora is not introduced in the test dataset because the results would be unreliable. Consequently, there are only two test datasets remaining. Test set pair 2 has 66 instances for the non-translated class and 36 instances for the translated class. Test set pair 3 has 28 non-translated class instances and 14 translated class instances.

Spanish Data						
Classifier	**Generic Data Representation**		**Excluding Simplification**		**Excluding Explicitation**	
	MTxt	TTxt	MTxt	TTxt	MTxt	TTxt
Baseline	64.71%	66.67%	64.71%	66.67%	64.71%	66.67%
Naïve Bayes	77.45%	**97.62%**	**78.43%**	85.71%	75.49%	95.24%
JRip	76.47%	92.86%	72.55%	80.95%	74.51%	95.24%
J48	73.53%	**97.62%**	71.57%	88.10%	71.57%	97.62%
IB1	77.45%	92.86%	69.61%	95.24%	76.47%	**100%**
SVM	78.43%	**97.62%**	77.45%	83.33%	78.43%	95.24%
Vote	**83.33%**	**97.62%**	**78.43%**	**100%**	**81.37%**	97.62%

Table 5.7: Classification Accuracy Results. Model trained on the entire dataset and evaluated on separate medical and technical test datasets.

In Table 5.7 the accuracies for the classifiers tested are reported. In the table, MTxt is a notation for the medical texts and TTxt for the technical ones. The first observation of the overall results, regardless of the learning model, is that the classifiers perform better on the technical domain compared to the medical domain.

For the generic model, the classifiers' performances vary between 92.86%, for JRip and IB1, and 97.62%, for all the rest of them. When excluding the simplification features, the classifiers achieve results between 80.95%, for JRip, and 100%, for the Vote meta-classifier. For the third learning model, the one which excludes the explicitation features, the classifiers vary much less, and they also yield the highest results of all: the accuracies range between 95.24% and 100%.

In general, the results decrease when excluding the features, except for the voting classifier on the technical data. It should also be noted that the technical domain has 42 instances in total, which may lead to higher

results compared to the medical one. Yet all the learning models appear to have consistently high results for this domain in particular.

For the medical domain, the results appear to be reasonably high when compared to the baseline. The Vote meta-classifier obtains the best results for all three learning models. It reaches up to 83.33% in the generic learning model, then the second best result, 81.37%, is for the excluding explicitation learning model, and finally, the excluding simplification learning model obtains an accuracy of 78.43%.

The next section presents the learning model which uses only the simplification features in its data representation.

5.2.3 Simplification Learning Model

The usage of simplification features only in the data representation constitutes the *simplification learning model*, a scenario presented in Section 4.3.3. As a remainder, the features considered in the Spanish experiments are the following: sentence length, sentence depth, proportion in texts of simple sentences, of complex sentences, and of zero sentences, word length, lexical richness, and information load.

Table 5.8 shows the accuracies obtained by each classifier. On 10-fold cross-validation, the highest accuracy is obtained by the Voting meta-classifier, reaching a performance of 78.22%, whilst the highest value on the test dataset is obtained by two learners: Naïve Bayes and Vote meta-learner, both obtaining 77.70% accuracy.

The variation amongst the learners being rather small, more precisely up to 4.66 on cross-validation and 3.38 on the test dataset; there are no

Spanish Data		
Classifier	*10-fold cross-validation*	*Test set*
Baseline	65.33%	64.86%
Naïve Bayes	74.22%	75%
JRip	78%	**77.70**%
J48	77.33%	77.03%
IB1	73.56%	74.32%
SVM	76.44%	75.68%
Vote	**78.22**%	**77.70**%

Table 5.8: Simplification Learning Model: Classification Accuracies. Evaluation mode: 10-fold cross-validation.

significantly worse learners. The Decision Tree obtains the lowest value for both cross-validation and test evaluation, having a performance of 74.22% on the former, and 75.68% on the latter. On test evaluation, another two classifiers obtain the same accuracy, namely the Naïve Bayes and SVM classifier.

To investigate to what extent the classifiers are able to identify each particular class, namely translated and non-translated, the corresponding precision and recall values are further discussed.

5.2.3.1 Precision, Recall and F-measure Values

In Table 5.9, the values for precision, recall and f-measure are reported for each learning algorithm used in the simplification learning model for a 10-fold cross-validation evaluation mode.

For the translated class, the best algorithm is the Vote meta-classifier, which obtained an f-score of 0.620. Although Vote does not have the highest precision or recall, both its components achieve high values for this class.

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.653	1	0.79	non-translated
Naïve Bayes	0.661	0.526	0.586	translated
	0.773	0.857	0.813	non-translated
JRip	**0.788**	0.5	0.612	translated
	0.778	**0.929**	**0.847**	non-translated
J48	0.745	0.526	0.617	translated
	0.782	0.905	0.839	non-translated
IB1	0.626	**0.590**	0.607	translated
	0.789	0.813	0.801	non-translated
SVM	0.727	0.513	0.602	translated
	0.776	0.898	0.833	non-translated
Vote	0.784	0.513	**0.620**	translated
	0.782	0.925	**0.847**	non-translated

Table 5.9: Simplification Learning Model: Precision, Recall and F-measure. Evaluation mode: 10-fold cross-validation.

The highest recall for the translated class is attained by the IB1 classifier, with a value of 0.590, being followed by the J48 and Naïve Bayes with a 0.526 recall score. In terms of precision, the JRip classifier obtains the highest score, reporting a 0.788 value. For the non-translated class, the highest f-measure score is attained by the Vote and JRip algorithms, being shortly followed by the J48 classifier. The Vote meta-classifier and JRip reach a score of 0.847, whereas the decision tree classifier obtains a value of 0.839. In terms of precision, the IB1 classifier has its highest value, 0.789, being closely followed by the Vote meta-learner and J48 classifier, which obtain the value of 0.782. The highest recall values for this class are 0.929 and 0.925 pertaining to the JRip classifier and Vote algorithm, respectively.

Overall, the learning model identifies the non-translated class better, regardless of the classifier used. This tendency would indicate that the model could benefit from the addition of some new features.

In the next paragraphs, the patterns retrieved by JRip and Decision Tree are outlined.

5.2.3.2 Translational Patterns

The pruned rule set retrieved by the JRip classifier is illustrated in Figure 5.7. The learning algorithm achieves 78% accuracy on 10-fold cross-validation evaluation, a similar result to the generic learning model for the same classifier, more precisely 78.22% accuracy. This indicates that the classifier relies heavily on the simplification features even for the generic learning model.

```
Rule 1: (lexicalRichness <= 0.16) and (sentenceLength >= 20.33)
=> class=translated (109.0/25.0)
Rule 2: => class=non_translated (341.0/72.0)
```

Figure 5.7: JRip output: Simplification Learning Model. Evaluation mode: 10-fold cross-validation.

The rule set has two rules: the first one uses lexical richness and sentence length for the translated class, whereas the second rule classifies the non-translated instances. Relying only on these two attributes, the classifier is able to distinguish between the two classes with a performance above the chance level.

The pruned output provided by the Decision Tree classifier, which obtains 77.33% accuracy on 10-fold cross-validation and a slightly lower

performance on test evaluation, more precisely 77.03%, is shown in Figure 5.8.

```
lexicalRichness <= 0.16
|   sentenceLength <= 16.81: non_translated (30.0)
|   sentenceLength > 16.81
|   |   complexSentences <= 0.56: translated (102.0/19.0)**
|   |   complexSentences > 0.56
|   |   |   infoLoad <= 0.64: non_translated (15.0/4.0)
|   |   |   infoLoad > 0.64: translated (2.0)
lexicalRichness > 0.16: non_translated (301.0/67.0)**
```

Figure 5.8: Simplification Learning Model: Pruned Decision Tree classifier output. Evaluation mode: 10-fold cross-validation.

The root of the decision tree is the lexical richness attribute, continuing to the next levels with sentence length attribute, indicating thus the two most relevant features for this classifier. For the next two levels, the complex sentences and information load attributes are considered.

It can be observed that the most used decision branch, detecting the translated class, classifies 102 instances, 19 out of which are misclassified. For the other class the root attribute, lexical richness, makes the decision for 301 instances, 67 out of which are incorrectly classified as non-translated.

The next paragraphs highlight the ranking of the features involved in the simplification learning model.

5.2.3.3 Feature Ranking

The results are reported in Table 5.10. As expected from the table which renders the feature ranking for the generic learning model, the top ranked feature is lexical richness.

Information Gain	**Chi squared**
lexicalRichness	lexicalRichness
sentenceLength	sentenceLength
wordLength	simpleSentences
simpleSentences	wordLength
zeroSentences	zeroSentences
infoLoad	infoLoad
sentenceDepth	*sentenceDepth*
ambiguity	*ambiguity*
complexSentences	*complexSentences*

Table 5.10: Simplification Learning Model: Attributes Ranking Filters.

It can be observed that the first two attributes, lexical richness and sentence length, are also considered as the most relevant by the J48 and JRip classifiers, just as both the ranking filters above. Largely, both filters rank the same features at approximately the same position.

The evaluation of the classifiers on the medical and the technical domains is reported in what follows.

5.2.3.4 Evaluation on Medical and Technical Domains

This additional experiment is conducted in order to assess to what extent the simplification learning model is able to categorise the texts for the medical and technical domain, separately. The accuracies of the classifiers are shown in Table 5.11.

The results indicate that the learning model is able to categorise the texts much better for the technical domain, whereas for the medical domain the results appear to reach up to only 72.55%. The highest results are obtained by Vote meta-classifier for the technical domain, and by JRip for the medical domain.

Spanish Data		
Classifier	**Simplification Learning Model**	
	MTxt	TTxt
Baseline	64.71%	66.67%
Naïve Bayes	69.61%	88.10%
JRip	**72.55**%	90.48%
J48	71.57%	92.86%
IB1	67.65%	90.48%
SVM	68.63%	92.86%
Vote	70.59%	**95.24**%

Table 5.11: Classification Accuracy Results. Model trained on the entire dataset and evaluated on separate medical and technical test datasets.

The results are unexpectedly high on the technical domain, taking into consideration the fact that the model uses only a few features: the lowest accuracy is achieved by Naïve Bayes, with 88.10% accuracy, whereas the highest value is attained by the Vote classifier, with 95.24% accuracy.

5.2.4 Explicitation Learning Model

The objective is to assess to what extent the learning model, having only the potential explicitation features in the training process, can distinguish between translated and non-translated texts.

Table 5.12 illustrates the results obtained for each classifier used. Probably the first aspect to note is that on 10-fold cross-validation evaluation, IB1 and SVM are outperformed by the baseline. Second, the rest of the classifiers appear to distinguish the classes approximately to the same extent as that to which the baseline does, showing limited improvement.

On the test evaluation, IB1 appears to achieve only 62.16% accuracy, being outperformed by the baseline. The rest of the classifiers seem to achieve slightly better values, reaching up to 71.62%.

Classifier	*10-fold cross-validation*	*Test set*
Baseline	65.33%	64.86%
Naïve Bayes	67.56%	68.24%
JRip	69.11%	71.62%
J48	65.56%	71.62%
IB1	58.67%	62.16%
SVM	64.44%	68.24%
Vote	68.67%	68.92%
Vote*	**69.33**%	**72.97**%

Table 5.12: Explicitation Learning Model: Accuracies for several classifiers.

Since the Majority Vote meta-classifier includes the vote of a classifier which has results below the baseline, namely the IB1 classifier, the Majority Vote between SVM, Naïve Bayes and JRip has been also reported. The results are reported in the same table, marking the new configuration of the meta-learner with *Vote**. Consequently, the results improved, obtaining 69.33% on 10-fold cross-validation and 72.97% on test evaluation.

Considering all the results, one possible explanation is that most of the learning algorithms applied on the explicitation learning model do not learn to distinguish between the translated and non-translated classes any better than the baseline itself. Next, the corresponding results in terms of precision, recall and f-measure for each class are reported in Table 5.13.

5.2.4.1 Precision, Recall and F-measure Values

As expected from the accuracy results outlined earlier, the detailed values for the classifiers are low. In general, for the translated class the results are very low, the f-measure score reaching up to 0.483, for the JRip classifier. The highest precision for this class is registered by the Vote classifier, whilst the highest recall is achieved by the IB1 classifier.

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.653	1	0.79	non-translated
Naïve Bayes	0.541	0.423	0.475	translated
	0.726	0.81	0.765	non-translated
JRip	0.575	0.417	**0.483**	translated
	0.73	0.837	0.78	non-translated
J48	0.506	0.288	0.367	translated
	0.693	0.85	0.763	non-translated
IB1	0.412	**0.449**	0.429	translated
	0.693	0.66	0.676	non-translated
SVM	0.409	0.058	0.101	translated
	0.657	0.956	0.778	non-translated
Vote	**0.800**	0.154	0.258	translated
	0.681	**0.979**	**0.803**	non-translated
Vote*	0.705	0.199	0.31	translated
	0.692	0.956	**0.803**	non-translated

Table 5.13: Precision, Recall and F-measure: Explicitation Learning Model. Evaluation mode: 10-fold cross-validation.

For the non-translated class, the highest values are attained as follows: the highest precision is obtained by JRip, the highest recall by Vote, and the highest f-measure by Vote and Vote*, both achieving a score of 0.803. The overall low accuracies seem to be caused by really low values for the

translated class. Therefore, considering only the explicitation features, the learning model appears to be unable to identify translated texts.

As the results reported for this learning model are very low, similar to the baseline, the translational patterns retrieved by the Decision Tree and the JRip classifiers are not reported (the patterns obtained are irrelevant).

Feature Ranking

Both algorithms, Information Gain and Chi-squared, rank the four attributes in the same order, namely: pronouns, adverbs, conjunctions, and sentences with one or more relative pronouns. Out of these, only the proportion of pronouns in a text seems to have had a score above zero when the filter evaluates using the entire training data. To analyse further the attributes, both ranking algorithms evaluate the features also using the 10-fold cross-validation method. Their results show the same ranking, only with values slightly above zero.

5.2.4.2 Evaluation on Medical and Technical Domains

Although the overall results show that the algorithms are not able to handle the categorisation task very well, the additional experiment on medical and technical domains is still conducted. The results are reported in Table 5.14.

Spanish Data		
Classifier	**Explicitation Learning Model**	
	MTxt	TTxt
Baseline	64.71%	66.67%
Naïve Bayes	61.76%	83.33%
JRip	63.73%	**90.48**%
J48	67.65%	80.95%
IB1	56.86%	76.19%
SVM	64.71%	66.67%
Vote	65.69%	80.95%
Vote*	**70.59**%	80.95%

Table 5.14: Classification Accuracy Results. Model trained on the entire dataset and evaluated on separate medical and technical test datasets.

The accuracies obtained are surprising as the classifiers show improved results on the technical domain, even when the learning model uses only

four attributes. Unexpectedly, the JRip classifier achieves 90.48% on the technical domain, being the highest accuracy obtained by the learning model. In contrast, the lowest result is achieved by SVM, being still unable to overtake the baseline. Nevertheless, Naïve Bayes, JRip, J48, and both Vote meta-classifiers attain remarkable results.

5.2.5 Ablation Study

The results of the ablation study are shown in Table 5.15. The values obtained by the classifiers for each feature available are evaluated using the 10-fold cross-validation evaluation mode. The baseline has the same value as before for the 10-fold cross-validation, more precisely 65.33%, according to the majority class considered in the training dataset.

According to the results obtained, most of the features do not improve compared to the baseline. This means that the classifiers are not able to categorise between translated and non-translated texts based solely on one attribute at a time. Although the generic learning model obtains remarkable results when it uses all the features available, the learners do not appear to handle the same task in a comparable manner in the ablation study.

Nevertheless, there is one attribute which appears to perform better than the baseline, namely lexical richness. The evaluation of this attribute is analysed below and the results obtained by the classifiers for the corresponding learning model are shown in Table 5.16.

Since IB1 performs below the baseline for the 10-fold cross-validation, a new configuration of the Vote meta-learner is adopted. For this experiment

Ablation Study		
Feature	**SVM**	**Vote**
Sentence Length	65.33%	66.44%
Sentence Depth	65.33%	65.11%
Zero Sentences	65.11%	64.67%
Simple Sentences	65.33%	65.56%
Complex Sentences	65.33%	65.11%
Ambiguity	65.33%	67.11%
Word Length	65.33%	67.11%
Lexical Richness	**69.11**%	**71.33**%
Information Load	65.56%	65.11%
Grammatical Words	65.33%	64%
Nouns	65.56%	66.67%
Finite Verbs	65.33%	66.22%
Adjectives	65.33%	65.78%
Adverbs	65.33%	65.33%
Numerals	63.78%	66%
Auxiliary Verbs	65.33%	65.11%
Pronouns	65.33%	63.78%
Prepositions	65.33%	65.33%
Determiners	65.33%	65.33%
Conjunctions	65.33%	65.33%
Gramm.W./Lex.W.	65.33%	65.56%
Sentences One or More Rel. Pron.	65.33%	66.44%

Table 5.15: Accuracy results for SVM and Majority Vote algorithms. Evaluation mode: 10-fold cross-validation.

only, the Majority Vote algorithm uses the output obtained from SVM, J48 and JRip classifiers and it is now marked with *Vote*** in the results table.

The highest results in each case are marked in bold: JRip achieves 72.22% accuracy on 10-fold cross-validation, whereas J48 obtains 72.30% on test dataset. The evaluation on the medical and technical tests is also included in this table, revealing interesting results.

It appears that the learning model which uses the highest ranked feature in the previous learning models, lexical richness, is able to categorise

Spanish Data				
Classifier	**Lexical Richness**			
	10-fold	*Test*	*On Domains*	
	Cross-validation	*Set*	*MTxt*	*TTxt*
Baseline	65.33%	64.86%	64.71%	66.67%
Naïve Bayes	69.56%	70.27%	65.69%	**78.57**%
JRip	**72.22**%	70.27%	68.63%	71.43%
J48	70.44%	**72.30**%	68.63%	**78.57**%
IB1	59.56%	48.65%	50%	% 45.24
SVM	69.11%	70.27%	65.69%	**78.57**%
Vote	71.33%	66.89%	**72.55**%	50%
Vote**	70.44%	70.27%	68.63%	71.43%

Table 5.16: Classification Accuracy Results. For the column marked 'On Domains' the learning model is trained on the entire dataset and is evaluated on separate medical and technical test datasets.

translated and non-translated texts in the technical domain with an unexpectedly high accuracy. Naïve Bayes, J48 and SVM classifiers achieve a performance of 78.57% on the classification task.

5.2.5.1 Translational Patterns

The translational patterns retrieved for the lexical richness learning model are shown in Figure 5.9. Both JRip and J48 classifiers are illustrated, presenting the patterns used in their classification. The former classifier obtains an accuracy of 72.22%, whilst the latter reports an accuracy of 70.44% using the patterns shown.

Although the performances obtained are modestly high, an accuracy above the baseline was expected to some extent for the top ranked feature of the generic learning model. For this reason, the fact that lexical richness is able to distinguish between translated and non-translated texts on its own is not a surprise. Nevertheless, the performance of the JRip classifier overtakes the baseline by classifier 6.89% accuracy.

```
JRIP rules:
------------------

Rule 1: (lexicalRichness <= 0.16) and (lexicalRichness <= 0.09)
=> class=translated (58.0/15.0)
Rule 2: (lexicalRichness <= 0.17) and (lexicalRichness >= 0.15)
and (lexicalRichness <= 0.16) => class=translated (53.0/20.0)
Rule 3: => class=non_translated (339.0/80.0)


J48 pruned tree:
------------------

lexicalRichness <= 0.16
|   lexicalRichness <= 0.08: translated (54.0/13.0)**
|   lexicalRichness > 0.08
|   |   lexicalRichness <= 0.14: non_translated (42.0/15.0)
|   |   lexicalRichness > 0.14: translated (53.0/20.0)
lexicalRichness > 0.16: non_translated (301.0/67.0)**
```

Figure 5.9: Learning Model for the Lexical Richness attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

In the next major section of this chapter, the experiments conducted on the Romanian data are presented.

5.3 Romanian Experiments

Similar to the Spanish experiments, the current section reports the results of the learning models keeping the same structure and the same evaluation modes (i.e., 10-fold cross validation and test set evaluation). The classifiers are trained by including and excluding the attributes proposed for the simplification or explicitation universal within the feature vector employed. The success rate indicates to what extent the model is influenced by the simplification features, and then t-test analyses whether this influence is statistically significant.

The translationese generic learning system exploits forty-seven multilingual features; these are all the features computed for the Romanian data. The reason for choosing this set of features as well as their argument to stand in favour of either one hypothesis or another was outlined in Chapter 4. Afterwards, simplification learning model and explicitation learning model are reported.

Throughout all the experiments, the training and test datasets have the same size. The data extracted from the corpus are divided into a training dataset and a separate test dataset. The training one comprises 639 instances: 223 for the translation class and 416 for non-translation class instances. The test dataset comprises 148 instances: 49 for the translation class and 99 for the non-translation class.

In the pre-processing stage of the analysis, before conducting all the experiments outlined in the previous chapter, the attributes are under investigation. To assess whether classifiers would perform better if the attributes were filtered first, the Chi-squared ranking algorithm is employed using the entire training dataset. The features are ranked and, then, the attributes ranked last, or more precisely those which obtained a score value of 0, are removed from the generic learning model. The results of the classifiers are evaluated.

The new model uses then the remaining thirty-five features and the classifiers achieve similar results as the model using all the forty-seven attributes available. This indicates that the classifiers are not improved by the filtering stage of the attributes, suggesting that even the last ranked attributes bring their contribution to the learning model. For this reason,

the generic model keeps all its original attributes in its configuration. The detailed results of the classifiers are evaluated using the 10-fold cross-validation method and are reported in Appendix C.1. The ranking of all the attributes available for Romanian as well as the generic learning model are reported shortly in the next section.

5.3.1 Translationese Generic Learning Model

The translationese generic learning model reports outstanding results and the accuracies of the classifiers are shown Table 5.17.

Romanian		
Classifier	*10-fold cross-validation*	*Test set*
Baseline	65.10%	66.89%
Naïve Bayes	96.09%	96.62%
JRip	93.90%	94.60%
J48	93.90%	96.62%
IB1	94.84%	100%
SVM	**98.90**%	98.65%
Vote	**98.90**%	**99.32**%

Table 5.17: Generic Learning Model: Classification Accuracies.

The SVM and Vote classifiers report the highest results on the 10-fold cross-validation, obtaining 98.90% accuracy, whilst the rest of the classifiers obtain values above 93.90%. On the test dataset evaluation, the classifiers obtain slightly higher results. The IB1 algorithm appears to reach 100% accuracy, whereas the others vary between 94.60% and 99.32% success rate.

At this point the first observation that arises is the fact that the IB1 classifier performs surprisingly well on the test dataset. Taking into

account that on the 10-fold cross-validation evaluation the algorithm does not indicate a similar high score and appears to have a similar success rate as the other classifiers, the value of the accuracy on the test dataset may be a result obtained by chance. It should also be noted that the test dataset comprises 148 instances, which may be seen as a small test dataset and can thus favour an unusual result of a classifier. For this reason, the second highest result is considered more reliable and marked in bold in the table.

Overall, the learning model has an outstanding performance indicating that the task of distinguishing between translated and non-translated texts can be achieved with accuracies between 93.90% and 98.90%.

5.3.1.1 Precision, Recall and F-measure Values

As expected from the results shown earlier, the corresponding detailed results by class reveal outstanding values. Table 5.18 reports the performance of the learning model in terms of precision, recall and f-measure score for each classifier.

The SVM and Vote algorithms appear to be the best learners for the partial results as well. They both attain the same results, misclassifying only 7 instances: 4 non-translated texts, and 3 translated ones. Note that the Vote meta-classifier considers first the output of the SVM in its decision making process, so it is expected that the Vote meta-learner to be highly influenced by it. They handle exceptionally well both translated and non-translated classes in terms of all the metrics: precision, recall, and consequently, the f-measure score. Although their results appear almost identical at this level, they do differ in terms of the mean error metrics

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.651	1	0.789	non-translated
Naïve Bayes	0.919	0.973	0.946	translated
	0.985	0.954	0.969	non-translated
JRip	0.934	0.888	0.910	translated
	0.941	0.966	0.954	non-translated
J48	0.922	0.901	0.912	translated
	0.948	0.959	0.953	non-translated
SVM	**0.982**	**0.987**	**0.984**	translated
	0.993	**0.990**	**0.992**	non-translated
IB1	0.899	0.960	0.928	translated
	0.978	0.942	0.960	non-translated
Vote	**0.982**	**0.987**	**0.984**	translated
	0.993	**0.990**	**0.992**	non-translated

Table 5.18: Generic Learning Model: Precision, Recall and F-measure for each Classifier. Evaluation mode: 10-fold cross-validation.

reported by the algorithms. These details for both learning algorithms are provided in Appendix C, in which Figure C.1 presents the summary of the SVM results, and Figure C.2 that of the Majority Vote meta-classifier.

The second best classifier, according to f-measure score, appears to be the Naïve Bayes learner, having a value of 0.969 for the non-translated class, and 0.946 for translated. Despite the overall high results, a slight tendency to handle the non-translated class better than the other class can be still observed.

In the next section, the translational patterns retrieved by the JRip and the J48 classifiers for the generic learning model are presented.

5.3.1.2 Translational Patterns

Highly recurring patterns characterising translational language and the use of potential features of simplification constitute a research gap in the literature. Such patterns are retrieved by JRip and decision tree classifiers.

The JRip rules obtained for the generic learning model are represented in Figure 5.10. The first rule is used to correctly identify 161 instances of the translated class, whereas the last rule classifies almost all of the non-translated instances available, namely 405 instances.

```
Rule 1: (InformationLoad <= 0.617284) and (LexicalRichness <= 0.493827)
 and (VbHasZPavg <= 0.381443) => class=translated (164.0/3.0)
Rule 2: (Nouns <= 0.303116) and (Adpositions >= 0.099026)
=> class=translated (56.0/8.0)
Rule 3: => class=non-translated (419.0/14.0)
```

Figure 5.10: Generic Learning Model: JRip classifier rules output. Evaluation mode: 10-fold cross-validation.

In terms of features, the JRip classifier relies firstly on information load, lexical richness and the proportion of verbs marked as having an AZP in their subject position. For its second rule, noun and adposition attributes are employed. All the other instances that do not fulfil the conditions imposed by the first two rules are classified using the third rule, being thus labelled as belonging to the translated class. Using these three rules, the learner obtains an accuracy of 93.90% on the 10-fold cross-validation evaluation and 94.60% on the test dataset.

Another classifier which provides the patterns retrieved in the classification process is Decision Tree. The pruned decision tree for the translationese generic learning model is represented in Figure 5.11.

```
InformationLoad <= 0.617934
|   pronPossessive <= 0.002342
|   |   LexicalRichness <= 0.507669: translated (206.0/8.0)**
|   |   LexicalRichness > 0.507669
|   |   |   VerbsMainSubjonctive <= 0.014458
|   |   |   |   Numerals <= 0.032099: non-translated (22.0)
|   |   |   |   Numerals > 0.032099
|   |   |   |   |   pronPersonal <= 0.005579: translated (6.0/1.0)
|   |   |   |   |   pronPersonal > 0.005579: non-translated (3.0)
|   |   |   VerbsMainSubjonctive > 0.014458: translated (4.0)
|   pronPossessive > 0.002342: non-translated (45.0)
InformationLoad > 0.617934
|   Nouns <= 0.3025
|   |   VerbsPersThreeSingular <= 0.022321: translated (9.0)
|   |   VerbsPersThreeSingular > 0.022321
|   |   |   Pronouns <= 0.05625
|   |   |   |   Adverbs <= 0.050536: translated (2.0)
|   |   |   |   Adverbs > 0.050536: non-translated (2.0)
|   |   |   Pronouns > 0.05625: non-translated (20.0)
|   Nouns > 0.3025: non-translated (320.0/5.0)**
```

Figure 5.11: Generic Learning Model: Pruned tree output from the Decision Tree classifier. Evaluation mode: 10-fold cross-validation.

The information load, a feature also assigned to simplification, appears to be the root of the decision tree for the translationese generic learning model. On the second level, nouns and possessive pronouns appear to be relevant for this classifier. The classifier also uses lexical richness and third person singular verbs, features which are placed on the next level of the tree. The use of the morphological sub-categories appears to have a beneficial outcome, the classifier being able to reveal interesting features that can be further investigated in qualitative studies.

The leaf nodes for each class are pointed out in the figure; for most of the translated class instances, information load, possessive pronouns and lexical richness are able to correctly identify 198 instances. For the other class, 315 instances are correctly identified using only the information load and the noun attributes. Using this decision tree, the algorithm achieves 93.90% accuracy on the 10-fold cross-validation and 96.62% accuracy on the test dataset.

5.3.1.3 Feature Ranking

The feature selection evaluators' outputs are further analysed in order to provide an overview of the most influential features of the learning model. The Information Gain and Chi-square algorithms provide the information reported in Figure 5.19. The table excludes the null-valued attributes, marking these in in italics in Appendix C.2.

As can be noticed, the ranking provided by both algorithms has approximately the same type of information. The same tendency was also observed in the Spanish experiments.

The first five features which most influence the classification are: information load, proportion of nouns, proportion of grammatical words per lexical words, proportion of prepositions, and lexical richness, two of which are considered to stand for the simplification universal. They are closely followed by another set of seven features: proportion of common nouns, the proportion of grammatical words, possessive pronouns, third singular verbs, numerals, verbs which have an AZP in the subject position, and complex sentences.

Regarding the simplification features investigated in these experiments, the ranking algorithms place three of them amongst the most influential features of the learning model: information load, which is also ranked first, followed by lexical richness, and the proportion of complex and simple sentences in texts. For the explicitation attributes, only the possessive pronouns attribute appears to be placed among the top ten ranked features, being followed by the verbs which have an AZP in the subject position attribute on the thirteenth position.

Romanian Data	
Information Gain	**Chi squared**
InformationLoad	InformationLoad
Nouns	Nouns
GrammaticalWperLexicalW	GrammaticalWperLexicalW
Adpositions	Adpositions
LexicalRichness	LexicalRichness
CommonNouns	CommonNouns
GrammaticalWords	GrammaticalWords
pronPossessive	pronPossessive
VerbsPersThreeSingular	VerbsPersThreeSingular
Numerals	Numerals
ComplexSentences	VbHasZPavg
SimpleSentences	SimpleSentences
VbHasZPavg	ComplexSentences
pronAdjDemonstrative	pronAdjDemonstrative
Determiners	VerbsMainIndicative
VerbsMainIndicative	Determiners
Conjunctions	Conjunctions
Adverbs	Adverbs
pronInterogRelative	ProperNouns
ProperNouns	pronInterogRelative
VerbsMainParticiple	VerbsMainParticiple
SentenceLength	VerbsMainGerund
VerbsMainGerund	pronPersonal
pronPersonal	SentenceLength
SentencesAtLeastOneRelPronoun	SentencesAtLeastOneRelPronoun
pronReflexive	pronReflexive
pronIndefinite	pronIndefinite
VerbsPersOnePlural	VerbsPersOnePlural
WordLength	WordLength
Pronouns	Pronouns
Verbs	Verbs
VerbsMainSubjonctive	VerbsPersTwoSingular
VerbsPersTwoSingular	VerbsAux
VerbsAux	VerbsMainSubjonctive
AdjectivesSuperlative	AdjectivesSuperlative
...	...

Table 5.19: Attributes Ranking Filters for the Translationese Generic Learning Model.

5.3.2 Comparison between Learning Models

Similarly to the Spanish experiments, two pairs of learning models are compared: the generic model with the excluding simplification learning model, and the generic model with the excluding explicitation learning model. The rationale is as follows: if the lack of simplification or explicitation features decreases the performance of the classifiers, then

this can be considered an argument for the existence of the corresponding hypothesis.

The performances of the learning algorithms for the 10-fold cross-validation evaluation on the training data and the accuracies obtained on the test dataset are reported in Table 5.20.

Romanian Data						
Classifier	**Generic Data Representation**		**Excluding Simplification Model**		**Excluding Explicitation Model**	
	10-fold cv.	*Test set*	*10-fold cv.*	*Test set*	*10-fold cv.*	*Test set*
Baseline	65.10%	66.89%	65.10%	66.89%	65.10%	66.89%
Naïve Bayes	96.09%	96.62%	94.37%	95.95%	94.68%	95.95%
JRip	93.90%	94.60%	93.58%	99.32%	93.11%	95.95%
J48	93.90%	96.62%	91.55%	97.30%	94.21%	97.30%
IB1	94.84%	100%	93.43%	100%	92.80%	100%
SVM	**98.90**%	98.65%	**97.18**%	97.97%	**98.44**%	97.30%
Vote	**98.90**%	**99.32**%	96.71%	**100**%	97.81%	**99.32**%

Table 5.20: Comparison between the learning models: Accuracies for several classifiers.

The baseline is 64.5% since the dominant class is the non-translated one, and all the learners outperform it.

At this point, a few observations emerge. First, on the 10-fold cross-validation evaluation, the best performance is consistently obtained by the SVM classifier throughout all the learning models employed. Second, on the test evaluation the IB1 classifier achieves 100% throughout all the learning models. This is an unexpected outcome, considering its values obtained for the 10-fold cross-validation evaluation. As pointed out earlier, this may occur by chance; maybe the test instances have high similarity with some instances from each class and since the test dataset is not very large this accuracy can appear. This is considered to be a chance result since in the

training data this classifier did not perform so well, being in line with the other classifiers.

Third, the results obtained by the classifiers are outstanding and they represent the highest success rate obtained amongst similar studies in the literature to date. The performance of the model constitutes proof that an automatic system is able to distinguish between translated and non-translated texts. Consequently, *evidence for the translationese hypothesis is thus brought forward in terms of the attributes considered in the data representation.*

Furthermore, the simplification and explicitation hypotheses are analysed. The removal of the simplification features leads to a slightly decreased accuracy for all the classifiers on the 10-fold cross-validation, having approximately 2-3% lower accuracies. T-test evaluation did not register any statistical difference between the translationese generic model and the excluding simplification learning model. This indicates that the model achieves comparable results with or without the simplification features. It should be noted though that the model aggregates several features, achieves very high results in both cases and, thus, may have enough attributes to be able to handle the task to the same extent even without a few attributes in its data representation.

Comparing the accuracies obtained on the test evaluation on the two learning models, a few classifiers achieve slightly better results on the excluding simplification learning model. These are: J48, JRip and the Vote meta-classifier. Nevertheless, considering that the results obtained by the 10-fold cross-validation are more reliable (because of a larger amount

of data tested), and that not all the classifiers have this tendency on the test dataset, the overall picture seems to indicate that *the exclusion of the simplification features leads to decreased performances of the classifiers.*

The similar tendency of having slightly lower values is noted also for the excluding explicitation learning model. However, there are two exceptions: first, on 10-fold cross-validation, for the learning model which excludes the explicitation features, the J48 classifier achieves a slightly lower value; second, the J48 and JRip are the exceptions for the test evaluation data.

More detailed results for each classifier are presented below: precision, recall and f-measure scores for each class are indicating to what extent the classifiers handle each category of text, translated and non-translated.

5.3.2.1 Excluding Simplification Learning Model

The first learning model reported is the model which excludes the simplification features from the data representation, being shortly followed by the excluding explicitation learning model in the next section.

Precision, Recall and F-measure Values

The accuracies of the classifiers for the former learning model are reported by class in terms of precision, recall and f-measure in Table 5.21.

The highest f-measure scores, for both the translated and non-translated classes, are reported for SVM: 0.956 for the translated class, and 0.976 for the non-translated class. The highest precision for the non-translated class is attained by Naïve Bayes, obtaining an outstanding value

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.651	1	0.789	non-translated
Naïve Bayes	0.882	**0.969**	0.923	translated
	0.982	0.93	0.956	non-translated
JRip	0.906	0.91	0.908	translated
	0.952	0.95	0.951	non-translated
J48	0.886	0.87	0.878	translated
	0.931	0.94	0.935	non-translated
IB1	0.872	0.951	0.91	translated
	0.972	0.925	0.948	non-translated
SVM	**0.956**	0.964	**0.96**	translated
	0.981	**0.976**	**0.978**	non-translated
Voting	0.943	0.964	0.953	translated
	0.981	0.969	0.975	non-translated

Table 5.21: Excluding Simplification Features Learning Model. Evaluation mode: 10-fold cross-validation.

of 0.982. The highest recall is attained by the SVM classifier, having a value of 0.976.

For the translated class, the results appear fairly similar, being only slightly lower than those reported for the non-translated class. The highest precision is obtained by the SVM classifier, 0.956, whereas the highest recall is attained by the Naïve Bayes classifier, more precisely 0.969.

Overall, the non-translated class appears to be slightly better identified by the learning algorithms, a tendency in line with the other learning models on the Spanish experiments.

Translational Patterns

This section presents the translational patterns retrieved for the learning model which excludes the simplification features in its data

representation. Figure 5.12 illustrates the rule set utilised by the JRip classifier.

```
Rule 1: (GrammaticalWperLexicalW >= 0.570292) and
(pronPossessive <= 0.002294) and (Nouns <= 0.297491)
and (Adverbs <= 0.054482) => class=translated (123.0/0.0)

Rule 2: (Adpositions >= 0.109462) and (Nouns <= 0.320463)
and (Adverbs <= 0.043909) => class=translated (46.0/0.0)

Rule 3: (Adpositions >= 0.115756) and (Nouns <= 0.334378) and
(pronPossessive <= 0) and (Numerals >= 0.03881)
=> class=translated (24.0/0.0)

Rule 4: (pronPossessive <= 0) and (Nouns <= 0.322281) and
(Determiners <= 0.032929) and (Adjectives >= 0.069374)
=> class=translated (21.0/3.0)

Rule 5: => class=non-translated (425.0/12.0)
```

Figure 5.12: Excluding Simplification Features: JRip classifier rules output. Evaluation mode: 10-fold cross-validation.

As reported in the figure, the JRip learner has an outstanding accuracy on the rules identifying the translated class. The first two rules are largely used in the identification of the translated class instances: the first one successfully classifies 123 instances, and the second one 46 instances. The aggregation of nouns, grammatical words per lexical words, possessive pronouns and nouns for the first rule, adpositions, nouns and adverbs for the second rule, as well as the features used in the third rule, appears to be highly accurate for the translated class: there are no instances misclassified as translated class.

Next, the decision tree obtained by the excluding simplification features model is illustrated in Figure 5.13. The grammatical words per lexical words attribute, as well as adpositions, appear on the first two levels of the tree.

```
GrammaticalWperLexicalW <= 0.562842
|   Adpositions <= 0.108209: non-translated (284.0/2.0)**
|   Adpositions > 0.108209
|   |   Nouns <= 0.305447: translated (13.0/1.0)
|   |   Nouns > 0.305447: non-translated (61.0/9.0)
GrammaticalWperLexicalW > 0.562842
|   pronPossessive <= 0.002299
|   |   Nouns <= 0.318436
|   |   |   Determiners <= 0.058041: translated (192.0/6.0)**
|   |   |   Determiners > 0.058041
|   |   |   |   Nouns <= 0.292011: translated (5.0/1.0)
|   |   |   |   Nouns > 0.292011: non-translated (6.0)
|   |   Nouns > 0.318436
|   |   |   Adpositions <= 0.133588: non-translated (24.0/3.0)
|   |   |   Adpositions > 0.133588: translated (8.0/1.0)
|   pronPossessive > 0.002299: non-translated (46.0)
```

Figure 5.13: Excluding Simplification Features: Pruned tree output from the Decision Tree classifier. Evaluation mode: 10-fold cross-validation.

In this decision tree, most of the translated instances are retrieved using the grammatical words per lexical words ratio and the possessive pronouns, determiners and nouns attributes; more precisely, 186 translated instances. For the other class, the majority of the instances, namely 282, are correctly identified using the grammatical words per lexical words ratio and the adpositions attributes.

In the next section, the results obtained for the other learning model, namely the excluding explicitation learning model, are reported.

5.3.2.2 Excluding Explicitation Learning Model

Overall, the learning model achieves 92.80% as the lowest accuracy of the 10-fold cross-validation evaluation, a value obtained by the IB1 classifier. The highest one, on the same type of evaluation, is attained by SVM with 98.44% accuracy.

As most of the classifiers, on this model, tend to lower their accuracy compared to the corresponding values for the generic learning model,

explicitation may have some support in its favour. Yet, the t-test evaluation did not indicate any statistical significance between the two models for any classifier used. Probably the dedicated learning model, namely the explicitation learning model, will provide more information in this regard. The explicitation learning model is further reported in the present chapter, in Section 5.3.4.

In the next paragraphs, the detailed results by class are reported for the learning model which excludes the explicitation features from its data representation.

Precision, Recall and F-measure Values

In Table 5.22 the results obtained using the 10-fold cross-validation evaluation are reported for each classifier utilised.

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.651	1	0.789	non-translated
Naïve Bayes	0.899	0.955	0.926	translated
	0.975	0.942	0.958	non-translated
JRip	0.909	0.892	0.900	translated
	0.943	0.952	0.947	non-translated
IB1	0.88	0.919	0.899	translated
	0.956	0.933	0.944	non-translated
J48	0.927	0.906	0.916	translated
	0.95	0.962	0.956	non-translated
SVM	**0.973**	**0.982**	**0.978**	translated
	0.990	**0.986**	**0.988**	non-translated
Vote	0.964	0.973	0.969	translated
	0.986	0.981	0.983	non-translated

Table 5.22: Excluding Explicitation Learning Model. Evaluation mode: 10-fold cross-validation.

From the results shown earlier, some outcomes, such as the results for the SVM classifier, are expected to have outstanding values. However, it is unexpected for one classifier to obtain all the highest results in terms of precision, recall, and f-measure for both classes involved, translated and not translated. This is the case for the SVM classifier in this learning model.

Furthermore, comparing the values obtained per class for each classifier employed, the non-translated class appears to have a slight edge over the other one. This behaviour is in line with the previous precision and recall results reported for the other two learning models, the generic model and the excluding simplification features model. Also, both classes report outstanding overall results.

Translational Patterns

Translational patterns in terms of a pruned decision tree output and a set of rules are illustrated in Figure 5.14 and 5.15, respectively.

```
InformationLoad <= 0.617934
|   Numerals <= 0.017007: non-translated (41.0/2.0)
|   Numerals > 0.017007
|   |   Adpositions <= 0.088652: non-translated (15.0)
|   |   Adpositions > 0.088652
|   |   |   LexicalRichness <= 0.527875
|   |   |   |   SentenceLength <= 34.136364: translated (202.0/4.0)**
|   |   |   |   SentenceLength > 34.136364
|   |   |   |   |   Interjections <= 0.000693
|   |   |   |   |   |   GrammaticalWords <= 0.362519: non-translated (6.0)
|   |   |   |   |   |   GrammaticalWords > 0.362519: translated (2.0)
|   |   |   |   |   Interjections > 0.000693: translated (3.0)
|   |   |   LexicalRichness > 0.527875
|   |   |   |   CommonNouns <= 0.244782: translated (2.0)
|   |   |   |   CommonNouns > 0.244782: non-translated (15.0)
InformationLoad > 0.617934
|   Nouns <= 0.3025
|   |   VerbsPersThreeSingular <= 0.022321: translated (9.0)
|   |   VerbsPersThreeSingular > 0.022321: non-translated (24.0/2.0)
|   Nouns > 0.3025: non-translated (320.0/5.0)**
```

Figure 5.14: Excluding Explicitation Learning Model: Pruned decision tree output. Evaluation mode: 10-fold cross-validation.

For the non-translated class, the decision tree classifier uses the information load and noun attributes for most of the instances. It correctly classifies 315 instances using only these two attributes. For the other class, more features are employed, namely: numerals, adpositions, lexical richness and sentence length. Note that three of these features are also marked as pertaining to the simplification hypothesis.

In Figure 5.15, the rule set retrieved by the JRip learning algorithm considers the following attributes: information load, lexical richness, simple sentences, and adpositions. Again, note that three out of these four features are indicators for the simplification hypothesis.

```
Rule 1: (InformationLoad <= 0.614665) and
(LexicalRichness <= 0.493827) and (SimpleSentences >= 0.703704)
and (Adpositions >= 0.08931) => class=translated (144.0/0.0)

Rule 2: (Nouns <= 0.302041) and (Adpositions >= 0.089905)
=> class=translated (70.0/9.0)

Rule 3: (Adpositions >= 0.111111) and (LexicalRichness <= 0.530885)
and (Adjectives >= 0.084195) and (AdjectivesComparative <= 0)
=> class=translated (14.0/3.0)

Rule 4:  => class=non-translated (411.0/7.0)
```

Figure 5.15: Excluding Explicitation Learning Model: JRip Rules. Evaluation mode: 10-fold cross-validation.

It should be also observed that the first rule identifies correctly 144 instances as belonging to translated class, having no misclassified instances of this type. The rule uses two of the most discussed attributes in the literature, lexical richness and information load, in combination with the simple sentences attribute and the adpositions feature.

5.3.3 Simplification Learning Model

As shown in Table 5.23, the results of the simplification learning model are lower compared to the generic model, but nevertheless remarkable for reaching a 94.68% accuracy on 10-fold cross-validation, and reporting values between 91.89% and 96.62% on the test dataset evaluation. The IB1 result is considered to be obtained by chance and, as a result, omitted in this discussion.

Classifier	*10-fold cross-validation*	*Test set*
Baseline	65.10%	66.89%
Naïve Bayes	92.64%	92.57%
JRip	92.18%	95.95%
J48	91.08%	93.92%
IB1	94.05%	100%
SVM	92.33%	91.89%
Vote	**94.68**%	**96.62**%

Table 5.23: Classification Accuracies: Simplification Learning Model.

The lower values of the classifiers' performances are expected, as this learning model uses only five features in its data representation compared to forty-seven in the generic learning model. Nevertheless, the classifiers learn the target concept with outstanding accuracies.

These results can be interpreted as an argument in favour of the simplification hypothesis since the classifiers handle the task with performances well above the chance level. These high values complement the overview provided by the research scenario in which the exclusion of the simplification features is assessed. The results emphasise the fact that

the simplification hypothesis appears to be validated in these experiments in terms of the features included in the data representation.

To analyse to what extent each class is detected in the learning model, the precision, recall and f-measure values are reported below.

5.3.3.1 Precision, Recall and F-measure Values

The corresponding results in terms of precision, recall and f-measure score for each class are reported in Table 5.24.

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.651	1	0.789	non-translated
Naïve Bayes	0.879	**0.915**	0.897	translated
	0.953	0.933	0.943	non-translated
JRip	0.902	0.87	0.886	translated
	0.932	0.95	0.94	non-translated
J48	0.892	0.848	0.869	translated
	0.92	0.945	0.932	non-translated
IB1	0.922	0.906	0.914	translated
	0.95	0.959	0.955	non-translated
SVM	0.892	0.888	0.89	translated
	0.940	0.942	0.941	non-translated
Vote	**0.935**	0.91	**0.923**	translated
	0.953	**0.966**	**0.959**	non-translated

Table 5.24: Simplification Learning Model. Evaluation mode: 10-fold cross-validation.

Overall, comparing all the f-measure scores obtained by the classifiers, the Vote meta-classifier appears to have the highest values for both classes: 0.923 for the translated class, and 0.959 for the non-translated class. It also achieves the highest results for almost all the other metrics, being overtaken

by Naïve Bayes in terms of recall score. For the non-translated class, Naïve Bayes obtains the same highest value for precision as the Vote algorithm.

5.3.3.2 Translational Patterns

Highly recurring patterns identifying one class or another are retrieved by the JRip and decision tree classifiers. These are further illustrated in Figure 5.16 and Figure 5.17, respectively. Both classifiers indicate that the information load attribute is a highly influential feature in the classification task.

```
Rule 1: (InformationLoad <= 0.616236) and (LexicalRichness <= 0.50495)
and (SimpleSentences >= 0.769231) => class=translated (135.0/0.0)

Rule 2: (InformationLoad <= 0.60356) and
(LexicalRichness <= 0.475548) => class=translated (52.0/6.0)

Rule 3: (InformationLoad <= 0.629382) and
(LexicalRichness <= 0.536705) and (SimpleSentences >= 0.708333)
and (WordLength >= 5.174758) => class=translated (22.0/2.0)

Rule 4: (InformationLoad <= 0.61324) and (LexicalRichness <= 0.527875)
and (SimpleSentences >= 0.619048) and (WordLength >= 4.882716)
=> class=translated (16.0/4.0)

Rule 5:  => class=non-translated (414.0/10.0)
```

Figure 5.16: Simplification Learning Model: JRip Rules. Evaluation mode: 10-fold cross-validation.

The first rule reported by JRip is the most used one, and combines the following features: information load, lexical richness, and simple sentences. These features are highly investigated in the literature in corpus-based studies, but none of them adopt machine learning techniques. The present model brings more rigorous evidence and it also indicates patterns formed

by these attributes in the learning process identifying translated and non-translated instances.

The decision tree classifier provides the pruned decision tree acquired for the simplification learning model in Figure 5.17.

```
InformationLoad <= 0.617934
|   LexicalRichness <= 0.50495
|   |   SimpleSentences <= 0.769231
|   |   |   InformationLoad <= 0.602804: translated (63.0/10.0)
|   |   |   InformationLoad > 0.602804: non-translated (22.0/7.0)
|   |   SimpleSentences > 0.769231: translated (135.0)**
|   LexicalRichness > 0.50495
|   |   SentenceLength <= 27.857143
|   |   |   LexicalRichness <= 0.528249
|   |   |   |   WordLength <= 4.994413: non-translated (5.0)
|   |   |   |   WordLength > 4.994413: translated (11.0/1.0)
|   |   |   LexicalRichness > 0.528249: non-translated (22.0/2.0)
|   |   SentenceLength > 27.857143: non-translated (28.0)
InformationLoad > 0.617934: non-translated (353.0/16.0)**
```

Figure 5.17: Simplification Learning Model: Pruned Decision Tree Output. Evaluation mode: 10-fold cross-validation.

In the decision tree outlined, most instances pertaining to the non-translated class are detected using only the information load feature. For the other class, the leaf node which correctly classifies most of the instances uses three levels of the decision tree, namely: information load, lexical richness and simple sentences attributes.

5.3.3.3 Feature Ranking

Both ranking filters provide largely the same output, placing information load, lexical richness and simple sentences in the first three positions. Afterwards, the complex sentences, sentence length and word length are

ranked. None of these attributes are marked with null-values in the evaluation on the full training set.

Chi-squared	**Information Gain**
InformationLoad	InformationLoad
LexicalRichness	LexicalRichness
SimpleSentences	SimpleSentences
ComplexSentences	ComplexSentences
SentenceLength	SentenceLength
WordLength	WordLength

Table 5.25: Feature Ranking for the Simplification Learning Model.

In the section below, the explicitation learning model is presented.

5.3.4 Explicitation Learning Model

In Table 5.26 the accuracies obtained for all the classifiers employed in the explicitation learning model are shown.

Classifier	*10-fold cross-validation*	*Test set*
Baseline	65.10%	66.89%
Naïve Bayes	**80.75**%	83.11%
JRip	79.34%	86.49%
J48	79.66%	82.43%
IB1	74.18%	100%
SVM	80.13%	79.73%
Vote	79.81%	**95.27**%

Table 5.26: Classification Accuracies: Explicitation Learning Model.

On the 10-fold cross-validation evaluation, IB1 seems to score the lowest accuracy, obtaining 74.18%. It has to be pointed out that the IB1 classifier seems to obtain 100% accuracy on the test set. Given that on the 10-fold cross-validation, the classifier obtains a much lower value, the results on the test set must appear by chance. On the test evaluation mode,

without considering the result obtained by IB1, the highest reliable result is obtained by Vote classifier, having 95.27% accuracy.

Overall, the results show that the learning model is able to handle the categorisation task with reliable accuracies, up to 80% for SVM and Naïve Bayes learning algorithms.

5.3.4.1 Precision, Recall and F-measure Values

The corresponding results by class for the explicitation learning model are reported in Table 5.27. These results are reported for the 10-fold cross-validation evaluation.

Detailed Results by Class				
Classifier	**Precision**	**Recall**	**F-Measure**	**Class**
Baseline	0	0	0	translated
	0.651	1	0.789	non-translated
Naïve Bayes	0.655	**0.946**	**0.774**	translated
	0.962	0.733	0.832	non-translated
JRip	0.695	0.726	0.711	translated
	0.850	0.829	0.839	non-translated
J48	0.700	0.731	0.715	translated
	0.852	0.832	0.842	non-translated
IB1	0.620	0.673	0.645	translated
	0.816	0.779	0.797	non-translated
SVM	**0.720**	0.704	0.712	translated
	0.843	**0.853**	**0.848**	non-translated
Vote	0.706	0.722	0.714	translated
	0.849	0.839	0.844	non-translated

Table 5.27: Explicitation Learning Model. Evaluation mode: 10-fold cross-validation.

The SVM classifier achieves the highest precision for the translated class, 0.720, and the highest recall and f-measure for the non-translated

class, namely 0.853 and 0.848, respectively. The remaining highest values, the highest precision for the non-translated class and the highest recall for the translated, are achieved by Naïve Bayes.

As a general tendency among the learners, the non-translated class appears to be better identified, which indicates that the explicitation learning model can benefit from the addition of more features which could be relevant for the translated class.

5.3.4.2 Translational Patterns

The rule set provided by the JRip classifier for the current learning model is shown in Figure 5.18.

```
Rule 1: (pronPossessive <= 0.001515) and (pronIndefinite >= 0.001259)
and (VbHasZPavg <= 0.353659) and (Pronouns >= 0.063462)
=> class=translated (94.0/11.0)

Rule 2: (pronPossessive <= 0.001757) and (VbHasZPavg <= 0.279279)
and (Pronouns >= 0.0623) and (pronReflexive <= 0.011844)
and (pronInterogRelative <= 0.015699)
=> class=translated (25.0/5.0)

Rule 3: (pronPossessive <= 0.001515) and (Pronouns >= 0.053836)
and (pronInterogRelative <= 0.007752) and
(Adverbs <= 0.050523) => class=translated (68.0/9.0)

Rule 4: => class=non-translated (452.0/61.0)
```

Figure 5.18: Explicitation Learning Model: JRip Rules. Evaluation mode: 10-fold cross-validation.

The output of the JRip learner indicates that the most relevant features in the learning process are: possessive pronouns, indefinite pronouns, AZP verbs and pronouns. To some extent, it is expected to note that the

verbs which have an AZP in the subject position attribute appear to be an influential feature in the classification task, given that it was ranked on the thirteen place among all the forty-seven features.

The second rule, used to a lesser extent compared to the first one, uses the possessive pronouns, the verbs which have an AZP in the subject position attribute, the pronouns attribute in general, the reflexive pronouns and the interrogative relative pronouns. The third rule includes the use of adverbs in the classification.

Next, the Decision Tree classifier reports a pruned tree having the possessive pronouns attribute at its root. This indicates that this feature is considered the most influential feature by this classifier. On the next two levels, the pronouns and conjunctions features are utilised in the classification.

```
pronPossessive <= 0.001757
|   Pronouns <= 0.052632: non-translated (73.0/16.0)
|   Pronouns > 0.052632
|   |   Conjunctions <= 0.068966: non-translated (37.0/7.0)
|   |   Conjunctions > 0.068966
|   |   |   pronInterogRelative <= 0.01084: translated (209.0/42.0)**
|   |   |   pronInterogRelative > 0.01084
|   |   |   |   pronIndefinite <= 0.001271: non-translated (36.0/9.0)
|   |   |   |   pronIndefinite > 0.001271
|   |   |   |   |   Pronouns <= 0.067454: non-translated (17.0/6.0)
|   |   |   |   |   Pronouns > 0.067454: translated (25.0/8.0)
pronPossessive > 0.001757: non-translated (242.0/1.0)**

Number of Leaves  :  7
Size of the tree :  13
```

Figure 5.19: Explicitation Learning Model: Pruned Decision Tree Output.

It can be observed that the features used by this classifier slightly differ from the rule set generated by the JRip learner in two ways: first, the AZP

verbs attribute as well as the adverbs and reflexive pronouns attributes are not being taken into consideration within the decision tree; second, the conjunctions attribute is preferred by the J48 learner and it places this feature on the third level of the tree, being part of the branch which identifies most of the translated class instances.

To achieve a better picture of the ranking of the features for this learning model, the same two evaluators are employed and their results are shown below.

5.3.4.3 Feature Ranking

The ranking of the attributes used in this model is reported in Table 5.28. Both Information Gain and Chi-squared algorithms fully agree regarding the order of the attributes.

Chi-squared	**Information Gain**
pronPossessive	pronPossessive
VbHasZPavg	VbHasZPavg
Conjunctions	Conjunctions
Adverbs	Adverbs
pronInterogRelative	pronInterogRelative
pronPersonal	pronPersonal
SentAtLeastOneIntRelPron	SentAtLeastOneIntRelPron
pronReflexive	pronReflexive
pronIndefinite	pronIndefinite
Pronouns	Pronouns
pronNegative	pronNegative

Table 5.28: Feature Ranking for the Explicitation Learning Model.

Possessive pronouns, AZP verbs, as well as the conjunctions attribute appear to be the top most reliable features obtained in the learning model. The first attribute was previously seen as being heavily utilised by both the decision tree algorithm and the JRip classifier.

The analysis of all the available attributes for the Romanian data continues by conducting the ablation study.

5.3.5 Ablation Study

The experiment employs a single feature at a time in the data representation of a learning model in order to investigate whether that feature on its own is able to distinguish between the two classes. The results obtained in the 10-fold cross-validation evaluation are presented in Table 5.29.

The baseline is the same as before, 65.10%, obtained by using the ZeroR classifier. Most of the features achieve similar accuracies as the baseline, meaning that the classifiers are not able to learn to distinguish reliably between the two classes.

It is important to emphasise that this outcome only indicates that they cannot perform the task of categorisation on their own, but it does not entail that those features are useless in this type of classification. They may appear to be useful in combination with other features, as is shown in the learning models presented earlier.

Nevertheless, there are a number of features that do handle the categorisation task beyond the chance level. These features and their results are presented in Table 5.30. Note that NB is an acronym for Naïve Bayes, and the best accuracy for each attribute is marked in bold in the table.

The accuracies obtained for the learning system are remarkable considering the fact that the model is using only one feature at a time. For these features, the ablation study reflects that the model is able to

Feature	SVM	Vote
Nouns	**84.51**%	**83.88**%
Verbs	65.10%	64.32%
Adjectives	65.10%	64.63%
Adverbs	65.10%	62.91%
Numerals	**74.80**%	**73.87**%
Pronouns	65.10%	64.95%
Adpositions	82.94%	**82.79**%
Determiners	64.79%	62.91%
Articles	65.10%	63.85%
Conjunctions	64.16%	64.86%
Gramm W. per Lexical W.	**82.63**%	**85.81**%
Grammatical Words	**78.40**%	**78.56**%
Interjections	64.95%	63.51%
Proper Nouns	65.10%	65.54%
Common Nouns	**80.59**%	**79.73**%
Vbs. 1st Pers. Pl.	65.10%	60.81%
Vbs. 1st Pers. Sg.	65.10%	65.54%
Vbs. 2nd Pers. Pl.	65.10%	66.89%
Vbs. 2nd Pers. Sg.	65.10%	66.89%
Vbs. 3rd Pers. Pl.	65.10%	64.86%
Vbs. 3rd Pers. Sg.	**76.21**%	**77.70**%
Aux. Vbs.	65.10%	66.89%
Modal Vbs.	65.10%	65.10%
Vbs. Indic.	65.10%	64.32%
Vbs. Subj.	65.10%	63.85%
Vbs. Imper.	65.10%	65.10%
Vbs. Inf.	65.10%	64.32%
Vbs. Ger.	65.10%	63.54%
Vbs. Part.	65.41%	63.85%
Compar. Adj.	65.10%	65.10%
Pos. Adj.	65.10%	64.95%
Superl. Adj.	65.10%	64.95%
Lexical Richness	**80.75**%	**81.53**%
Sentence Length	65.26%	64.01%
Word Length	65.10%	62.91%
Simple Sent.	**72.77**%	**74.49**%
Complex Sent.	**72.77**%	**75.43**%
Information Load	**83.72**%	**83.88**%
Vbs.AZP	**71.99**%	**71.67**%
Sent. one or more int-rel. pron.	64.63%	62.75%
Interrogative Relative Pron.	65.10%	62.60%
Personal Pron.	65.10%	65.41%
Negative Pron.	65.10%	65.26%
Reflexive Pron.	65.10%	64.63%
Possessive Pron.	**74.18**%	**70.11**%
Indefinite Pron.	65.10%	62.28%
Demonstrative Pron. and Adj.	65.57%	64.48%

Table 5.29: Accuracy results for the Ablation Study. Evaluation mode: 10-fold cross-validation.

reliably perform the categorisation task even when using one feature in its data representation.

Most of the attributes report the highest results for the SVM and Voting algorithms, but there are a few which also have high values for

Feature	NB	JRip	J48	IB1	SVM	Vote
Nouns	84.19%	84.04%	82.16%	79.97%	**84.51**%	83.88%
Numerals	74.18%	73.71%	72.14%	63.22%	**74.80**%	73.87%
Gramm W. per Lexical W.	82.94%	82.63%	83.10%	75.59%	82.63%	**85.81**%
Grammatical Words	78.25%	78.25%	77.78 %	74.02%	78.40%	**78.56**%
Adpositions	82.47%	82.63%	82.63%	75.27%	**82.94**%	82.79%
Common Nouns	**80.75**%	79.19%	78.87%	70.89%	80.59%	79.73%
Vbs. 3rd Pers. Sg.	75.74%	76.53%	76.06%	69.33%	76.21%	**77.70**%
Lexical Richness	81.22%	**82.16**%	78.72%	71.52%	80.75%	81.53%
Simple Sent.	73.40%	72.93%	72.61%	72.30%	72.77%	**74.49**%
Complex Sent.	73.24%	72.46%	72.61%	72.30%	72.77%	**75.43**%
Information Load	84.35%	84.35%	**84.98**%	78.09%	83.72%	83.88%
Vbs. AZP	72.14%	**72.46**%	72.14%	64.01%	71.99%	71.67%
Possessive Pron.	**74.18**%	72.77%	72.77%	71.36%	**74.18**%	70.11%

Table 5.30: 10-fold Cross-validation Evaluation on Particular Features: Several Classification Results.

Naïve Bayes or JRip. Out of all the attributes marked in this table, the features which are also potential features of simplification are the following, listed in the order of accuracy: information load, lexical richness, complex sentences and simple sentences.

Similarly, for the potential explicitation features, the list comprises two features, irrespective of how low their accuracy is: possessive pronouns and verbs which have AZP in their subject position.

The ablation study also yields unexpected results, considering the high accuracy obtained for the ratio of grammatical words per lexical words, nouns, adpositions and common nouns. Perhaps this may indicate that the difference in their usage throughout translated and non-translated texts appears as a consequence of redundancy, which may help in the categorisation task.

The translational patterns retrieved by these features are shown in the next section.

5.3.5.1 Translational Patterns

The patterns provided for the features which obtained greater values than 80% accuracy for either one of these two classifiers are shown below, whereas the outputs for the remaining features are illustrated in Appendix C.4.

The learning model using information load in its data representation has the highest values for the JRip and J48 classifiers. Their patterns are shown in Figure 5.20.

```
JRIP rules:
------------------

Rule 1:(InformationLoad <= 0.615551)
=> class=translated (269.0/69.0)
Rule 2: => class=non-translated (370.0/23.0)

J48 pruned tree:
------------------
InformationLoad <= 0.617934: translated (286.0/79.0)**
InformationLoad > 0.617934: non-translated (353.0/16.0)**
```

Figure 5.20: Learning Model for the Information Load attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

With the patterns retrieved in the learning process, the JRip learning algorithm obtains an accuracy of 84.35%, whilst the decision tree classifier achieves even a slightly better value, namely 84.98%, the latter one being also the highest performance achieved among all the classifiers.

The second best results for the two classifiers are obtained for the learning model which uses the nouns attribute. This is an unexpected outcome. Nevertheless, it is a reasonable one since the literature points out several investigations on lexical richness and information load. These

two attributes involve the use of the main morphological classes, obviously including nouns.

The outputs retrieved for the nouns learning model are illustrated in Figure 5.21.

```
JRIP rules:
------------------

Rule 1: (Nouns <= 0.303716) => class=translated (222.0/49.0)
Rule 2: => class=non-translated (417.0/50.0)

J48 pruned tree:
------------------

Nouns <= 0.318261
|   Nouns <= 0.292994: translated (160.0/24.0)**
|   Nouns > 0.292994: non-translated (149.0/70.0)
Nouns > 0.318261: non-translated (330.0/17.0)**
```

Figure 5.21: Learning Model for the Nouns attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

It is remarkable to obtain such an accuracy for the learning model that uses only the nouns attribute: the JRip classifier obtains 84.04% accuracy, a value close to the success rate reported by SVM, whilst the J48 classifier achieves a performance of 82.16% with the decision tree reported.

The next best performance is obtained by the learning model that uses the grammatical words per lexical words attribute. The patterns retrieved for the JRip and J48 classifiers are illustrated in Figure 5.22.

JRip classifier achieves 82.63% accuracy using the reported set of rules, whereas the J48 learning model obtains a similar performance having 83.10% success rate. The findings retrieved by this experiment are surprising because there are no studies in the literature investigating

```
JRIP rules:
------------------

Rule 1: (GrammaticalWperLexicalW >= 0.570225)
=> class=translated (267.0/73.0)
Rule 2: => class=non-translated (372.0/29.0)

J48 pruned tree:
------------------

GrammaticalWperLexicalW <= 0.562842: non-translated (358.0/23.0)**
GrammaticalWperLexicalW > 0.562842: translated (281.0/81.0)**
```

Figure 5.22: Learning Model for the Grammatical Words per Lexical Words Attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

this feature. To the best of the author's knowledge, this work is the first to analyse this characteristic and also to provide this kind of information about it.

The next learning model with best accuracies for these two classifiers which provide patterns is the model which uses the adpositions attribute. Its output is shown in Figure 5.23.

```
JRIP rules:
------------------

Rule 1: (Adpositions >= 0.108844) and (Adpositions >= 0.126853)
=> class=translated (126.0/15.0)
Rule 2: (Adpositions >= 0.109347) => class=translated (133.0/57.0)
Rule 3: => class=non-translated (380.0/36.0)

J48 pruned tree:
------------------

Adpositions <= 0.108642: non-translated (379.0/35.0)**
Adpositions > 0.108642: translated (260.0/72.0)**
```

Figure 5.23: Learning Model for the Adpositions attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

Using the rules and the pruned decision tree reported, both learning algorithms are able to categorise between translated and non-translated texts with a performance of 82.63% accuracy. The adpositions feature was ranked four in the generic learning model among the most influential features in the classification. Yet, it is surprising to obtain such accuracy for this experiment.

The fifth best performance in terms of JRip and J48 classifiers is achieved by the learning model which uses lexical richness in its data representation. Figure 5.24 shows the JRip rule set retrieved in the learning process as well as the pruned decision tree obtained by the J48 classifier.

```
JRIP rules:
------------------

Rule 1: (LexicalRichness <= 0.491139)
=> class=translated (266.0/78.0)
Rule 2: => class=non-translated (373.0/35.0)

J48 pruned tree:
------------------

LexicalRichness <= 0.506349: translated (325.0/115.0)**
LexicalRichness > 0.506349: non-translated (314.0/13.0)**
```

Figure 5.24: Learning Model for the Lexical Richness attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

Lexical richness is a widely investigated feature in the research on translational hypotheses. Its presence among the features whose learning model achieves high accuracy on its own is expected to some extent; it is not surprising, mainly because of the good results obtained in the ablation study on the Spanish experiments. In the current ablation study, its

accuracy is higher, obtaining 82.16% accuracy for JRip classifier and 78.72% for the decision tree learning algorithm.

The patterns for the rest of the learning models for the remaining features are detailed in Appendix C.4. Overall, the ranking of the features reported earlier, for the generic learning model, places most of these attributes among the top most influential features of the model and, thus, performance above the baseline might have been expected to some extent.

In the next section, a discussion of all the experiments presented for Spanish and Romanian data is outlined.

5.4 Discussion and General Remarks

The learning models presented in this chapter come as a result of each objective outlined in Chapter 1. Note that the most important tables with results are Table 5.4, for Spanish, and Table 5.20, for Romanian, because the accuracies obtained indicate that the generic translationese learning model performs better than the models which exclude the simplification features and explicitation features, respectively. These results thus bring evidence in favour of the simplification and explicitation hypotheses. An overview of each set of experiments is provided in the following paragraphs.

The main outcomes obtained for the Spanish experiments, according to each translational hypothesis, are pointed out. The *translationese* hypothesis appears to be confirmed by the results, the learning model implemented being able to automatically categorise between translated and non-translated texts with very high accuracies.

Simplification, in terms of the features considered, appears to be confirmed by the learning models, its features appearing to be ranked amongst the first in the top relevant features of the model. In the comparison stage, the t-test indicates a statistical improvement when these features are included, and moreover, the learning model implemented for simplification, comprising only potential simplification features, achieves results varying between 73% and 78%. These accuracies show that the learning model is able to learn the target concept.

In contrast, for the explicitation, the results do not clearly indicate that the model reliably learns to distinguish between translated and non-translated texts using the explicitation features. Although in the comparison scenario, the t-test appears to be statistically significant for the SVM and Vote meta-classifier, indicating that the addition of the explicitation features had a statistically improved accuracy, the results of the explicitation learning model only reach up to 69.33% accuracy.

Considering these two outcomes, the *explicitation* features appear to be modestly reliable when they are combined with other sets of features. The additional experiment, evaluating the classifiers on medical and technical domains separately, indicates an interesting and intriguing twist: for the medical domain, the explicitation features achieve modest results, whereas for the technical results, some classifiers reach up to 80%-90% accuracy. This indicates that the domain on which the learning model is evaluated can influence its overall performance.

For the *Romanian* experiments, the learning models benefit from a wider range of attributes. This occurs in order to prove that the

translationese learning model can be *adaptable*, from language to language, using the natural language processing tools available according to the particular language under investigation.

The outcomes of the learning models implemented for Romanian show that the *translationese and simplification* hypotheses, in terms of the features under examination, are validated. The translationese generic learning model achieves accuracies above 93.90%, reaching up to 98.90% on 10-fold cross-validation, and even higher for the test evaluation. The simplification learning model also achieves outstanding results, its performance varying between 91.08% and 94.68% on 10-fold cross-validation, and even higher on the test data.

The explicitation learning model on Romanian reports lower results than its generic learning model, but still remarkably high: the classifiers' performances range between 74.18% and 80.75%. Despite the fact that the t-test did not register any statistical performance when the explicitation features were included, the explicitation learning model alone shows that it is able to learn the target concept relying on these potential features. As a conclusion, the findings indicate that *explicitation* appears to have an argument in its favour.

It should be noted that in the Romanian experiments, in the comparison scenario, given the number of features employed, many of them reliable in the learning process, the addition of the explicitation features may be overshadowed and, thus, there is no statistical significance to the result as an outcome. The ablation study indicates that *lexical richness* is able to achieve remarkable results on its own, for both Romanian and

Spanish. Moreover, on Romanian, a set of twelve features appear to handle the same task with good results. This indicates that even though simplification and explicitation appear to be highly plausible, they do not provide a comprehensive account of the nature of translational language.

The *patterns* reported throughout the thesis bring to light the interaction of some expected features, and some unexpected ones, paving the way for future hypotheses that may be supported by more rigorous evidence. This is achieved because of the adoption of the machine learning and natural language processing tools in the research process: from the extraction of the features to the retrieval of the patterns.

Also, the fact has to be emphasised that the patterns retrieved in this thesis should be treated as recurring patterns highly likely to appear in the context of the present experiments, and not as general or law-like tendencies. In this direction, the additional experiment emphasises the unexpected element when the explicitation learning model achieves surprisingly high accuracies in the technical domain. This may be an indication that the tendencies noted in translation, in this case explicitation, may not be universal since they appear to be *domain-dependent at this point.*

To hypothesise, translationese *may* universally occur in translational language, but *this does not necessarily imply that there is a set of universal features*, regardless of language or context. What may be suggested to be universal is the tendency in general, whereas its sub-hypotheses may be restricted by certain conditions (e.g., domain, genre, language, target readers, etc.). Maybe there is no universal in terms of a predefined feature

or set of features, but there may be a universal in the sense of a *universal distinctiveness* between translation and non-translation.

In the next section, the closest research studies relevant to this thesis are pointed out.

5.4.1 Comparison to Related Work

These experiments have been explicitly corroborated and strengthened by research studies within the Machine Translation area (Koppel and Ordan, 2011; Lembersky et al., 2011; Volansky et al., 2011; Lembersky et al., 2012). This is because the language models built on translated data appear to pave the way to higher performances of their SMT frameworks. The main distinction between these studies is that their research questions focus on the interference hypothesis, and do not necessarily restrict their model to being multilingual. Features involving n-grams are frequently used in their research, an aspect which implies that their system is language-dependent.

Another important research study relevant to the present work is the one reported by Baroni and Bernardini (2006). Their research question is similar, in that they investigate the extent to which a computer is able to detect translated texts, but their means differ. They also use and combine n-gram features, implying that their research study is language-dependent. Moreover, they only use the SVM classifier. In addition, the present research addresses more than just that research question.

To the best of my knowledge, the present research is novel in terms of the methodology employed for the investigation of simplification and explicitation. It is the first study to report translational patterns and

rankings of the features considered relevant in the task of distinguishing the translated texts from the non-translated ones. It is also the first study to address two languages at the same time, using an extensive set of features, on distinct domains.

5.4.2 Strengths and Limitations of the Learning Models

The main strength of the present research is its ability to learn the target concept with remarkable accuracy. It aggregates a large amount of features, being able to extract the most relevant ones and rank them. In this way, the patterns do not need to be handcrafted, as they can be retrieved using well-known learning algorithms.

The limitation of this study is mainly represented by the lack of the accuracy of the features involved. With automatic extraction, a degree of noise would be incurred in the pre-processing stage. Nonetheless, this limitation is taken into account by the learning algorithms which are able to discard irrelevant features in their classification.

Another limitation of the current approach represents its qualitative side: several linguistic phenomena and their explanations may be explained by them. Natural language processing tools are not always able to capture all the linguistic bits of the texts, or to capture them correctly. This means that the present work can be continued in this direction as further research, taking into consideration the features highlighted by the present implementations.

Still the reasons for adopting this approach are well-grounded, and its outcomes show that it is worth applying machine learning techniques within this field, and adopting it for further research analysis within the translation studies area.

5.5 Conclusions

This chapter reports the experiments implemented for the investigation of translational language, and the two related sub-hypotheses relevant to this thesis: simplification and explicitation. The outcomes of the five research scenarios justified in Chapter 4 are reported, and their results are detailed.

The first part of the chapter reports the Spanish experiments in Section 5.2, whilst the second part outlines the Romanian experiments in Section 5.3. Both these sections have the same structure, reporting the following learning models and research settings: the translationese generic learning model, the comparison of the generic model with the excluding simplification learning model, the comparison of the generic model with the excluding explicitation learning model, the detailed results of the excluding simplification learning model, the detailed results of the excluding explicitation learning model, the simplification learning model, the explicitation learning model, and the ablation study.

After the Romanian experiments are reported and their findings detailed, Section 5.4 discusses the overall outcomes and interprets the overall results in terms of translationese, simplification and explicitation. The same section also includes the comparison of the current research to the relevant related work, and points out its main strengths and limitations.

The next chapter provides an overview of the entire thesis, revisiting its main aims and objectives, and also pointing out to what extent these have been accomplished.

Chapter 6

Conclusions

6.1 General Conclusions

This thesis reports research in the area of Descriptive Translation Studies. The major aim of the research is to investigate the nature of translational language and its translationese manifestations, focusing on two important hypotheses previously defined within the domain: simplification and explicitation.

This thesis proposes a novel methodology for investigating these hypotheses: it investigates the possibility of using machine learning techniques relying both on features specific to the investigated hypotheses, namely simplification-specific features and explicitation-specific features, and on morphological features assumed to be generally characteristic of translationese. To this end, a learning framework is designed and implemented for the identification of translational language. The framework is modelled to solve a categorisation task, the goal of the learning

algorithms being to distinguish between translated texts and non-translated texts.

The second and third main aims of this research are the retrieval of the recurring patterns that are revealed in the process of solving the categorisation task, and the ranking of the most influential characteristics used to accomplish the learning task.

These aims are fulfilled through the implementation of a system that adopts the machine learning methodology proposed by this research. Moreover, the learning framework proves to be an adaptable multilingual framework for the investigation of the nature of translational language, its adaptability being illustrated in this thesis by its application to the investigation of two languages: Spanish and Romanian.

The present research is an interdisciplinary investigation which is situated at the confluence of three areas: Descriptive Translation Studies, on the one hand, and two sub-fields of Artificial Intelligence: Machine Learning and Natural Language Processing, on the other hand. The present thesis combines these three areas for the following reasons. First, the major aim of the study is to investigate the nature of translational language, placing the current research principally in the translation studies domain. Second, the chosen approach is machine learning because of its ability to handle a large set of features at the same time, to extract patterns, and to point out correlations between the target concept and the features under investigation. The third reason arises from the need to automatically extract values for a large set of features that are considered to be relevant

in the learning process. Consequently, existing natural language processing tools are used for the automatic extraction of values for these features.

The choice of machine learning as the appropriate methodology for the investigation of translational hypotheses is the fundamental innovation of this research: the interdisciplinary study possesses important potential for gaining further insights into this research topic. The present thesis proves how this approach can be modelled for the investigation of translationese, simplification and explicitation, and justifies its necessity within the domain of translation studies.

The fundamental *rationale* of the present learning framework is as follows: if translational language differs from non-translational language according to the features 'speculated' in the existing literature, then these features can be employed in the categorisation task that distinguishes between the two types of text.

The *findings of this research* show that machine learning algorithms relying on a set of features largely discussed in the literature are able to differentiate translated texts from non-translated ones with outstanding accuracies, providing thus a rigorous methodology for the investigation of various translational hypotheses.

These outcomes underline the existence of certain recurrent patterns identified in the current context, interpreted as recurring patterns with a high likelihood of being relevant in the separation of translated texts from non-translated ones, and not as absolute laws of translational language.

It is important to emphasise that this research did not aim at retrieving the universal patterns, nor the universal features of translational language,

but at extracting the patterns that occur with a high probability in the given context, for the type of corpus used for Spanish and Romanian.

This research shows that machine learning models, aggregating the different types of feature described in previous literature and in this work, are able to handle the intrinsic multi-dimensionality of the problem, at the same time enhancing the perspective on translational language. Moreover, the patterns retrieved in the learning process can pave the way for novel insights into translationese and existing translational hypotheses, and lead to a more refined theoretical background of the domain.

6.2 Aims and Contributions Revisited

The introductory chapter outlines the aims, the research questions and the objectives of the present research. This section summarises how these aims have been fulfilled by this research:

Objective 1 was to provide an overview of related research on the investigation of translational language, an objective that was achieved in Chapter 2. The strengths and drawbacks of existing work were pointed out.

Objective 2 was to identify and propose a suitable methodology for the investigation of translational language and its related translational hypotheses. The methodology had to be easily adaptable to distinct languages in order to enable the investigation of the nature of translational language, regardless of source or target languages. Such a multilingual methodology was proposed in Chapter 2, Section 2.4.

Objective 3 was to describe the data required for the methodology proposed. Comparable corpora comprising translated and non-translated texts were considered to be the most appropriate type of corpus for this investigation. To highlight the multilingual nature of the proposed methodology, two languages were investigated in this work: Spanish and Romanian. While Spanish comparable corpora of this type already existed within the domain of translation studies, a Romanian comparable corpus comprising such types of texts was not readily available. For this reason, the compilation of the necessary Romanian corpus was undertaken to provide the necessary resources for this research, and the corresponding details were provided in Chapter 3.

Objective 4 was to acquire the necessary tools for the proposed methodology. Given that the methodology adopted by this research implied the automatic extraction of features from naturally occurring text, a set of natural language processing tools were employed. A description of these tools was provided in Chapter 3.

Objective 5 was to propose a methodology for the investigation of potentially discriminative features characterising translational language, focusing on the research questions outlined in the introductory chapter. This objective is achieved in Chapter 4. The design of a learning model for the investigation of translationese was reported.

Objective 5.1 was to design the learning framework, which involved the selection of the features to be investigated. This objective

was accomplished in Chapter 4 through the design of the research settings and research scenarios, by pursuing the research questions, and by analysing the potential features of translational language using a multilingual framework. The configuration of the translationese generic learning model was thus designed, together with another two research scenarios that helped assess the impact of the simplification and explicitation features on the generic model.

Objective 5.2 was to design the simplification learning model: the learning model used to investigate the potential features of simplification. This objective was addressed in Chapter 4.

Objective 5.3 was to design the learning model for analysing the potential features of explicitation, namely the explicitation learning model. This objective was achieved in Chapter 4.

Objective 6 was to implement and evaluate all the learning models designed for the investigation of translational language. Chapter 5 fulfilled this goal by providing details on the implementation of the learning framework, as well as by reporting the evaluation results.

Objective 6.1 was to implement, evaluate and analyse the outcomes of the translationese generic learning model, and it was extensively described in Chapter 5.

Objective 6.2 was to implement the corresponding learning models and analyse to what extent simplification features were influencing the translationese generic learning model. This objective was addressed in Chapter 5.

Objective 6.3 was to implement the corresponding learning models and analyse to what extent explicitation features were influencing the translationese generic learning model. This objective was fulfilled in Chapter 5.

Objective 6.4 was to implement and analyse the findings of the simplification learning model. This objective was achieved in Chapter 5.

Objective 6.5 was to implement and analyse the findings of the explicitation learning model. This objective was accomplished in Chapter 5.

Objective 6.6 was to identify and discuss the strengths and limitations of the methodology adopted in this thesis. The objective was fulfilled in Chapter 5.

Objective 7 was to provide further directions of research, an objective that is addressed in the present chapter.

6.3 Review of the Thesis

The present thesis comprises six chapters, in which the objectives of the research are followed systematically. This section presents a general overview of the thesis by summarising each chapter.

Chapter 1 introduced the reader to the research topic investigated in this thesis, highlighting the aims, objectives and research questions set to be addressed, together with the original contributions made by this work.

Chapter 2 performed a comprehensive review of existing work on translationese and various translational hypotheses, focusing on the hypotheses that are most relevant to this work: simplification and explicitation. This chapter also outlined the proposed direction of research undertaken in this thesis.

Chapter 3 presented the resources and tools employed in this work. Given that the focus of this research was to propose a general methodology for the investigation of translationese and translational hypotheses that could be applied to texts in any target language, and texts translated from any source language, the investigation focused on two different languages: Spanish and Romanian. The comparable corpora comprising translated and non-translated texts in these two languages were described in this chapter. As this type of resource was not available for Romanian, the process of compiling the Romanian comparable corpus was also described in this chapter. General notions related to the discipline of machine learning, as well as Weka, the tool employed for performing the experiments, were presented in the second part of this chapter.

Chapter 4 described the methodology adopted in this research to investigate potentially discriminative features of translational language. The design of a machine learning framework that explores the nature of translational language by modelling a categorisation task between translated and non-translated texts was reported in this chapter. This framework relies on different types of feature grouped under three main learning models: the translationese generic learning model, the simplification learning model and the explicitation learning model. These features were described in detail in this chapter.

Chapter 5 provided details of the implementation of all the learning models presented, and reported the evaluation results for Spanish and Romanian in five research scenarios that rely on different combinations of features. The results and findings of each research scenario were analysed and discussed for each language. The strengths and shortcomings of the current methodology were highlighted at the end of this chapter.

In the present chapter, **Chapter 6**, conclusions have been drawn, and the aims and objectives of the thesis have been revisited, along with an estimate of the extent to which they have been accomplished in the experiments conducted. In addition, future directions of research are identified in the following section.

6.4 Further Directions of Research

Further threads of research can be extended from this work, involving both the translation studies area and the natural language processing area.

For the translation studies area, the learning models may be adapted for other languages, according to the availability of the corresponding natural language processing tools specific to that language. In this thesis, translationese and two of its related hypotheses are investigated. Nevertheless, the learning framework can be extended to investigate more hypotheses of the domain. Aggregating more features assumed to stand for a hypothesis H within the translationese generic learning model, and then evaluating to what extent the learning model is influenced by the new features, can be one thread of further research. Similar research can

be conducted thus for interference, convergence, etc., to reveal how the combination of the features interact with each other and which feature appears to be the most influential in the learning framework.

Another line of research would be to enhance the research on translation by embedding typical features obtained from the influence of other disciplines since the translation domain itself is characterised to be an interdisciplinary field. The flexibility of the machine learning approach, and the advantage that NLP tools provide, offer a methodological background for further lines of investigation from distinct perspectives.

Thus, collaborative research between scholars with interests in cognitive studies, cultural studies, translation technology, statistics, data mining, as well as computational linguistics is probably a promising ground for the investigation of the true nature of translational behaviour and its main characteristics.

Another thread of research which can be pursued from the starting point provided by the current learning framework would be the inclusion of a qualitative analysis of the translational patterns revealed in the process of solving the categorisation task. Further work stemming from the patterns extracted and the ranking of the features obtained can refine the actual hypotheses and thus advance the domain of translation studies.

Besides the research on the translation domain, other fields that can benefit from this research are located within the natural language processing domain itself: namely, the learning framework can be integrated into a system for the automatic compilation of parallel corpora by retrieving the candidate parallel texts. In a cross-plagiarism detection framework,

the current research can be adapted and integrated to identify whether the potentially plagiarised paragraph is in fact a translation. For the SMT systems, especially those handling a large amount of data, the learning models developed in this work can be integrated as a filtering module which assesses whether a text or a web page is already a translation, written either by humans or by automatic machine translation tools. If so, the SMT frameworks may choose not to consider the given text as relevant, thus reducing the errors in their statistically significant patterns.

Appendix A

Previously Published Work

Related experiments and some of the work reported in the thesis have been previously published in peer-reviewed international conference proceedings, journals and books. This research has been improved and adapted to the context of the thesis. The papers are shortly described below and their relatedness to this work is pointed out:

- **Ilisei, I.**, Inkpen, D., Corpas, G., Mitkov, R. (2012) *Romanian Translational Corpora: Building Comparable Corpora for Translation Studies*, in Proceedings of the Fifth Workshop on Building and Using Comparable Corpora, co-located with Language Resources and Evaluation International Conference (LREC2012), Istanbul, Turkey, pp. 56-61.

  This publication provides the details of the compilation of translational corpora for Romanian, RoTC. It comprises translated and non-translated texts, and the entire process is described in this paper. The compilation details from this paper are included in Chapter 3, as the RoTC corpus has been used for the Romanian experiments.

- **Ilisei, I.**, Mihăilă, C., Inkpen, D., and Mitkov, R. (2011) *The Impact of Zero Pronominal Anaphora on Translational Language: A Study on Romanian Newspapers*, in Proceedings of the International

Conference on Knowledge Engineering, Principles and Techniques, KEPT2011, Cluj-Napoca, Romania, July 46, pp. 43-50.

The research in this publication presents a different learning model than the ones reported in this thesis: it combines simplification and explicitation features within its data representation and it utilises distinct learning algorithms. As the publication also includes an additional analysis, namely the learning model that relies only on the anaphoric zero pronoun, this experiment is integrated within the ablation study for the Romanian experiments but the classifiers are using distinct settings in the learning model.

- **Ilisei, I.** and Inkpen, D. (2011) *Translationese Traits in Romanian Newspapers: A Machine Learning Approach*, in International Journal of Computational Linguistics and Applications, vol. 2, no. 1-2.

  This publication reports similar experiments as the research scenario which compares the translationese generic learning model and the model which excludes the simplification features. As the translationese generic model for Romanian language has been enriched for the current thesis, and distinct classifiers have been used, the results obtained differ from the earlier publication.

- **Ilisei, I.**, Inkpen, D., Corpas, G., and Mitkov, R. (2010) *Identification of Translationese: A Supervised Learning Approach*, in A. Gelbukh (Ed.): Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science Series 6008, Springer Berlin/Heidelberg, pp. 503-511.

  This paper reports similar experiments for Spanish data, describing the research scenario which compares the translationese generic learning model and the model which excludes the simplification features. However, the present thesis reports enriched data representation for the learning system and utilises distinct algorithms and data, and as a result, the findings obtain differ from this publication.

- Mihăilă, C., **Ilisei, I.**, and Inkpen, D. (2010) *To Be or Not To Be A Zero Pronoun: A Machine Learning Approach for Romanian*, in

Tufiş, D. and Forăscu, C. (Eds.) Multilinguality and Interoperability in Language Processing with Emphasis on Romanian, Romanian Academy Publishing House, pp. 303-316.

This paper reports the identification stage of the anaphoric zero pronouns for Romanian language, and it was build in order to include this feature within the learning models reported in Chapter 4. The feature appears to have an important role in the learning models presented in this thesis.

- Mihăilă, C., **Ilisei, I.**, and Inkpen, D. (2010) *Romanian Zero Pronoun Distribution: A Comparative Study*, in Proceedings of the Seventh International Conference on Language Resources and Evaluation - LREC 2010, Malta, European Language Resources Association (ELRA), 2010, 144-148.

  This publication presents early research on the anaphoric zero pronouns in order to build the environment for the identification stage of this feature. This research was conducted to automatically retrieve this feature and include it in the learning models for the Romanian experiments.

- **Ilisei, I.**, Inkpen, D., Corpas, G., and Mitkov, R. (2009) *Towards Simplification: A Supervised Learning Approach*, in Proceedings of Machine Translation Twenty-Five Years On, London, United Kingdom.

  This paper reports the impact of the simplification features on the generic learning model for Spanish. This is early research in the same direction as this thesis, however, the experiments reported in this present thesis are on distinct data, being also improved by enriching the overall model, the features involved and the classifiers used.

Appendix B

Spanish Experiments

B.1 Filtering the Attributes

Spanish Data	
Classifier	*10-fold cross-validation*
Baseline	65.33%
Naïve Bayes	75.56%
JRip	76.89%
J48	80%
IB1	77.11%
SVM	78.44%
Voting	80.89%

Table B.1: Spanish Generic Learning Model after filtering the attributes: Classification Accuracies.

B.2 Translationese Generic Learning Model: Feature Ranking

Spanish Data	
Information Gain	**Chi-squared**
lexicalRichness	lexicalRichness
finiteVerbs	finiteVerbs
numerals	numerals
adjectives	adjectives
sentenceLength	sentenceLength
prons	prons
simpleSentences	wordLength
wordLength	simpleSentences
grammaticalWords	zeroSentences
zeroSentences	nouns
nouns	infoLoad
infoLoad	grammaticalWords
ambiguity	*complexSentences*
complexSentences	*ambiguity*
sentAtLeastOneIntRelPron	*sentenceDepth*
grammsWPerLexicsWords	*grammsWPerLexicsWords*
sentenceDepth	*sentAtLeastOneIntRelPron*
auxVerbs	*auxVerbs*
adverbs	*adverbs*
conjs	*conjs*
preps	*preps*
dets	*dets*

Table B.2: Attributes Ranking Filters for the Generic Learning Model.

Appendix C

Romanian Experiments

C.1 Filtering the Attributes

Romanian Model	
Classifier	*10-fold cross-validation*
Naïve Bayes	96.2441%
JRip	93.8967%
J48	94.0532%
IB1	94.2097%
SVM	**98.9045**%
Vote	98.5915%

Table C.1: Romanian Generic Learning Model after filtering the attributes: Classification Accuracies.

C.2 Translationese Generic Learning Model: Feature Ranking

Romanian Data	
Information Gain	**Chi squared**
InformationLoad	InformationLoad
Nouns	Nouns
GrammaticalWperLexicalW	GrammaticalWperLexicalW
Adpositions	Adpositions
LexicalRichness	LexicalRichness
CommonNouns	CommonNouns
GrammaticalWords	GrammaticalWords
pronPossessive	pronPossessive
VerbsPersThreeSingular	VerbsPersThreeSingular
Numerals	Numerals
ComplexSentences	VbHasZPavg
SimpleSentences	SimpleSentences
VbHasZPavg	ComplexSentences
pronAdjDemonstrative	pronAdjDemonstrative
Determiners	VerbsMainIndicative
VerbsMainIndicative	Determiners
Conjunctions	Conjunctions
Adverbs	Adverbs
pronInterrogRelative	ProperNouns
ProperNouns	pronInterrogRelative
VerbsMainParticiple	VerbsMainParticiple
SentenceLength	VerbsMainGerund
VerbsMainGerund	pronPersonal
pronPersonal	SentenceLength
SentencesAtLeastOneRelPronoun	SentencesAtLeastOneRelPronoun
pronReflexive	pronReflexive
pronIndefinite	pronIndefinite
VerbsPersOnePlural	VerbsPersOnePlural
WordLength	WordLength
Pronouns	Pronouns
Verbs	Verbs
VerbsMainSubjonctive	VerbsPersTwoSingular
VerbsPersTwoSingular	VerbsAux
VerbsAux	VerbsMainSubjonctive
AdjectivesSuperlative	AdjectivesSuperlative
pronNegative	*pronNegative*
AdjectivesComparative	*Adjectives*
AdjectivesPositive	*AdjectivesPositive*
VerbsMainInfinitive	*AdjectivesComparative*
Adjectives	*VerbsPersThreePlural*
VerbsPersThreePlural	*Interjections*
VerbsPersTwoPlural	*VerbsPersTwoPlural*
VerbsPersOneSingular	*VerbsPersOneSingular*
Articles	*VerbsMainImperative*
VerbsMainImperative	*VerbsMainInfinitive*
Interjections	*VerbsModal*
VerbsModal	*Articles*

Table C.2: Attributes Ranking Filters for the Translationese Generic Learning Model.

C.3 SVM and Vote Classifiers for the Generic Learning Model

```
Correctly Classified Instances         632               98.9045 %
Incorrectly Classified Instances         7                1.0955 %
Kappa statistic                          0.9759
Mean absolute error                      0.0264
Root mean squared error                  0.1026
Relative absolute error                  5.8185 %
Root relative squared error             21.5331 %
Total Number of Instances              639

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.987     0.01       0.982     0.987     0.984      0.995    translated
                 0.99      0.013      0.993     0.99      0.992      0.995    non-translated
Weighted Avg.    0.989     0.012      0.989     0.989     0.989      0.995

=== Confusion Matrix ===

   a   b   <-- classified as
 220   3 |   a = translated
   4 412 |   b = non-translated
```

Figure C.1: Summary Results for the SVM classifier. Generic Learning Model using 10-fold cross-validation evaluation.

```
Correctly Classified Instances         632               98.9045 %
Incorrectly Classified Instances         7                1.0955 %
Kappa statistic                          0.9759
Mean absolute error                      0.011
Root mean squared error                  0.1047
Relative absolute error                  2.41   %
Root relative squared error             21.9578 %
Total Number of Instances              639

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.987     0.01       0.982     0.987     0.984      0.988    translated
                 0.99      0.013      0.993     0.99      0.992      0.988    non-translated
Weighted Avg.    0.989     0.012      0.989     0.989     0.989      0.988

=== Confusion Matrix ===

   a   b   <-- classified as
 220   3 |   a = translated
   4 412 |   b = non-translated
```

Figure C.2: Summary Results for the Vote meta-classifier. Generic Learning Model using 10-fold cross-validation evaluation.

C.4 Ablation Study: Translational Patterns

```
JRIP rules:
------------------

Rule 1: (VerbsPersThreeSingular <= 0.022508)
=> class=translated (220.0/71.0)
Rule 2: => class=non-translated (419.0/74.0)

J48 pruned tree:
------------------

VerbsPersThreeSingular <= 0.022544: translated (221.0/71.0)**
VerbsPersThreeSingular > 0.022544: non-translated (418.0/73.0)**
```

Figure C.3: Learning Model for the Third Person Singular Verbs attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

```
JRIP rules:
------------------

Rule 1: (pronPossessive <= 0.001515)
=> class=translated (382.0/164.0)
Rule 2: => class=non-translated (257.0/5.0)

J48 pruned tree:
------------------

pronPossessive <= 0: translated (358.0/150.0)**
pronPossessive > 0
|   pronPossessive <= 0.001247: non-translated (11.0)
|   pronPossessive > 0.001247
|   |   pronPossessive <= 0.001515: translated (13.0/3.0)
|   |   pronPossessive > 0.001515: non-translated (257.0/5.0)**
```

Figure C.4: Learning Model for the Possessive Pronouns attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

```
JRIP rules:
------------------

Rule 1: (VbHasZPavg <= 0.278351)
=> class=translated (173.0/61.0)
Rule 2: (VbHasZPavg <= 0.372881) and
(VbHasZPavg >= 0.366337) => class=translated (20.0/6.0)
Rule 3: => class=non-translated (446.0/97.0)

J48 pruned tree:
------------------

VbHasZPavg <= 0.278351: translated (173.0/61.0)
VbHasZPavg > 0.278351: non-translated (466.0/111.0)
```

Figure C.5: Learning Model for the Verbs which have an AZP in the Subject Position attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

```
JRIP rules:
------------------

Rule 1: (ComplexSentences <= 0.217391)
=> class=translated (241.0/93.0)
Rule 2: => class=non-translated (398.0/75.0)

J48 pruned tree:
------------------

ComplexSentences <= 0.227273: translated (264.0/107.0)
ComplexSentences > 0.227273: non-translated (375.0/66.0)
```

Figure C.6: Learning Model for the Complex Sentences attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

```
JRIP rules:
------------------

Rule 1: (SimpleSentences >= 0.791667)
=> class=translated (228.0/86.0)
Rule 2: => class=non-translated (411.0/81.0)

J48 pruned tree:
------------------

SimpleSentences <= 0.771429: non-translated (375.0/66.0)
SimpleSentences > 0.771429: translated (264.0/107.0)
```

Figure C.7: Learning Model for the Simple Sentences attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

```
JRIP rules:
------------------

Rule 1: (GrammaticalWords >= 0.346084) and
(GrammaticalWords >= 0.366022) => class=translated (143.0/31.0)
Rule 2: (GrammaticalWords >= 0.344894) and
(GrammaticalWords <= 0.363462) and (GrammaticalWords >= 0.356282)
=> class=translated (48.0/15.0)
Rule 3:  => class=non-translated (448.0/78.0)

J48 pruned tree:
------------------

GrammaticalWords <= 0.344423: non-translated (332.0/28.0)
GrammaticalWords > 0.344423: translated (307.0/112.0)
```

Figure C.8: Learning Model for the Grammatical Words attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

```
JRIP rules:
------------------

Rule 1: (CommonNouns <= 0.255988)
=> class=translated (258.0/79.0)
Rule 2: => class=non-translated (381.0/44.0)

J48 pruned tree:
------------------

CommonNouns <= 0.255988: translated (258.0/79.0)
CommonNouns > 0.255988: non-translated (381.0/44.0)
```

Figure C.9: Learning Model for the Common Nouns attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

```
JRIP rules:
------------------

Rule 1: (Numerals >= 0.03155) => class=translated (230.0/83.0)
Rule 2: => class=non-translated (409.0/76.0)

J48 pruned tree:
------------------

Numerals <= 0.02623: non-translated (326.0/44.0)
Numerals > 0.02623
|   Numerals <= 0.039474: non-translated (171.0/76.0)
|   Numerals > 0.039474: translated (142.0/39.0)
```

Figure C.10: Learning Model for the Numerals attribute: Translational patterns provided by the JRip and J48 classifiers. Evaluation mode: 10-fold cross-validation.

Appendix D

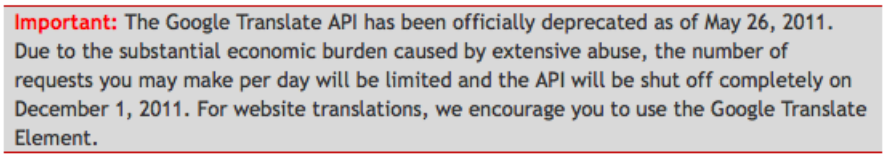
Translation API Issue

Article published on 12 June 2011 in *The Atlantic*[1] written by Fallows (2011).

Article title: "**An 'Economic Burden' Google Can No Longer Bear?**"

> "This is insider-tech talk, but I think it is very interesting in its implications – about language, "big data," Google's strategies, and the never-ending recalibration of goods vs bads, "signal to noise," on the internet.
>
> [Brief summary of what follows: Google is dropping an automatic-translation tool, because overuse by spam-bloggers is flooding the internet with sloppily translated text, which in turn is making computerized translation even sloppier.]
>
> There has been a rumble in the tech world about Google's announcement last month that it was "deprecating," and phasing out, its "Translate API." In simplest terms that means that website developers will no longer be able to use code that makes Google's translation algorithms automatically provide material for other sites. The standalone Google Translate site, which allows you to enter text or URLs for translation,

[1] Links: `http://www.theatlantic.com/technology/archive/2011/06/an-economic-burden-google-can-no-longer-bear/240283/`



will remain (along with some other features that apply Google translations to others' sites). But as an announcement on the Translate API site said:

For a very, very detailed explication of what this "economic burden" might mean for Google, check this analysis from the eMpTy Pages site on translation technology and related topics. Here is the part of the explanation that, for me, had the marvelous quality of being obvious – once it's pointed out – and interesting too:

The intriguing problem is the way that over-use of automatic translation can make it harder for automatic translation ever to improve, and may even be making it worse. As people in the business understand, computerized translation relies heavily on sheer statistical correlation. You take a huge chunk of text in one language; you compare it with a counterpart text in a different language; and you see which words and phrases match up. The computer doesn't have to "understand" either language for this to work. It just notices that the English words "good" or "goods" show up as bon in French in certain uses (i.e., as in "opposite of bad"), but as a variety of other French words depending on the context in English – "dry goods," "I've got the goods," "good grief," etc.

Crucially, this process depends on "big data" for its improvement. The more Rosetta stone-like side-by-side passages the system can compare, the more refined and reliable the correlations will become. Day by day and comparison by comparison, the translation will only get better. So that some day, in principle, we could understand anything written in any language, without knowing that language ourselves.

UNLESS ... the side-by-side texts used to "train" the system aren't any more accurate and nuanced than what the computer already knows. That is the problem with a rapidly increasing volume of machine-translated material. These computerized translations are better than nothing, but at best they are pretty rough. Try it for yourself: Go to the People's Daily Chinese-language home site; plug any story's URL (for instance, this one) into the Google Translate site; and see how closely the result resembles real English. You will get the point of the story, barely. Moreover, since these side-by-side versions reflect the computerized-system's current level of skill, by definition they offer no opportunity for improvement.

That's the problem. The more of this auto-translated material floods onto the world's websites, the smaller the proportion of good translations the computers can learn from. In engineering terms, the signal-to-noise ratio is getting worse. It's getting worse faster in part because of the popularity of Google's Translate API, which allows spam-bloggers and SEO operations to slap up the auto-translated material in large quantities. This is the computer-world equivalent of sloppy overuse of antibiotics creating new strains of drug-resistant bacteria. (Or GIGO – Garbage In, Garbage Out – as reader Rick Jones mentioned.) As the eMpTy Pages analysis describes the problem, using another analogy (emphasis added):

"Polluting Its Own Drinking Water ...An increasing amount of the website data that Google has been gathering has been translated from one language to another using Google's own Translate API. Often, this data has been published online with no human editing or quality checking, and is then represented as high-quality local language content....

It is not easy to determine if local language content has been translated by machine or by humans or perhaps whether it is in its original authored language. By crawling and processing local language web content that has been published without any

human proof reading after being translated using the Google Translate API, Google is in reality "polluting its own drinking water."...

The increasing amount of "polluted drinking water" is becoming more statistically relevant. Over time, instead of improving each time more machine learning data is added, the opposite can occur. Errors in the original translation of web content can result in good statistical patterns becoming less relevant, and bad patterns becoming more statistically relevant. Poor translations are feeding back into the learning system, creating software that repeats previous mistakes and can even exaggerate them."

That's all I have about this story, which I offer because it reveals a problem I hadn't thought of – and illustrates one more under-anticipated turn in the evolution of the info age. The very tools that were supposed to melt away language barriers may, because of the realities of human nature (i.e., blog spam) and the intricacies of language, actually be re-erecting some of those barriers. For the foreseeable future, it's still worth learning other languages."(Article published on 12 June 2011 in *The Atlantic* by James Fallows)

References

Aha, D. and Kibler, D. (1991), 'Instance-based learning algorithms', *Machine Learning* **6**, pp. 37–66.

Atkins, S., Clear, J. and Ostler, N. (1992), 'Corpus Design Criteria', *Literary and Linguistic Computing* **7**(1), pp. 1–16.

Baker, M. (1992), *In Other Words*, London: Routledge.

Baker, M. (1993), *Text and Technology: In Honour of John Sinclair*, Amsterdam: John Benjamins Publishing Company, chapter Corpus Linguistics and Translation Studies Implications and Applications, pp. 233–252.

Baker, M. (1995), 'Corpora in Translation Studies. An Overview and Suggestions for Future Research', *Target* **7**(2), pp. 223–243.

Baker, M. (1996), *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, Amsterdam: John Benjamins Publishing Company, chapter Corpus-based Translation Studies: The Challenges that Lie Ahead, pp. 175–186.

Baker, M. (1999), 'The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators', *International Journal of Corpus Linguistics* **4**, pp. 281–298.

Baker, M. (2004), 'A corpus-based view of similarity and difference in translation', *International Journal of Corpus Linguistics* **9**(2), pp. 167–193.

Baker, M. (2007), 'Patterns of Idiomaticity in Translated vs. Non-translated Text', *Belgian Journal of Linguistics* **21**, pp. 11–21.

Balaskó, M. (2008), 'What does the figure show? Patterns of translationese in a Hungarian comparable corpus', *Trans-Kom* **1**(1), pp. 58–73.

Balázs, M. (2011), *Terminology and Translation Studies. Plurilingual Terminology in the Context of European Intercultural Dialogue*, Cluj-Napoca: Scientia Publishing House, chapter Explicitation Strategies In the Translation of Emily Brontës Wuthering Heights, pp. 305–315.

Baroni, M. and Bernardini, S. (2003), A preliminary analysis of collocational differences in monolingual comparable corpora, In: D. Archer, P. Rayson, A. Wilson and T. McEnery, eds, *Proceedings of Corpus Linguistics 2003*, Lancaster: UCREL, pp. 82–91.

Baroni, M. and Bernardini, S. (2006), 'A new approach to the study of translationese: Machine-learning the difference between original and translated text', *Literary and Linguistic Computing* **21**(3), pp. 259–274.

Barrón-Cedeño, A., Potthast, M., Rosso, P. and Stein, B. (2010), Corpus and Evaluation Measures for Automatic Plagiarism Detection, In: *Proceedings of the LREC 2010 Conference*, pp. 771–774.

Becher, V. (2009), The Explicit Marking of Contingency Relations in English and German Texts: A Contrastive Analysis, In: *Societas Linguistica Europaea - 42nd Annual Meeting, Workshop: Connectives Across Languages*, University of Lisbon.

Becher, V. (2011*a*), Explicitation and Implicitation in Translation. A Corpus-based Study of English-German and German-English Translations of Business Texts, PhD thesis, University of Hamburg.

Becher, V. (2011*b*), 'When and why do translators add connectives? A corpus-based study', *Target* **23**(1), pp. 26–47.

Beretta, M. (1982), 'Problemi testuali della traduzione: casi di ambiguità anaforica in alice nel paese delle meraviglie', *Linguistica contrastiva. Atti del Convegno Internazional di Studi, Asti 26-28 maggio 1979, Società Linguistica Italiana* **20**, pp. 229–254.

Bernardini, S. and Zanettin, F. (2004), *Translation Universals. Do They Exist?*, Amsterdam: John Benjamins Publishing Company, chapter When is a Universal not a Universal?, pp. 51–62.

Biber, D. (1993), 'Representativeness in Corpus Design', *Literary and Linguistic Computing* **8**(4), pp. 243–257.

Blum-Kulka, S. (1986), *Interlingual and Intercultural Communication Discourse and Cognition in Translation and Second Language Acquisition Studies*, Vol. 35, Tubingen: Gunter Narr Verlag, chapter Shifts of Cohesion and Coherence in Translation, pp. 17–35.

Blum-Kulka, S. and Levenston, E. (1983), *Strategies in Interlanguage Communication*, Longman, chapter Universals of Lexical Simplification, pp. 119–139.

Borin, L. and Prütz, K. (2001), 'Through a Glass Darkly: Part-of-speech Distribution in Original and Translated Text', *Language and Computers. Amsterdam: Rodopi* **37**(1), pp. 30–44.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. and Scuse, D. a. (2012), WEKA Manual for Version 3-7-6, Technical report, The University of Waikato.

Brants, T., Popat, A., Xu, P., Och, F. and Dean, J. (2007), Large language models in machine translation, In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague*, pp. 858–867.

Bublitz, W. (1998), *Handbook of Pragmatics*, Amsterdam: John Benjamins Publishing Company, chapter Cohesion and Coherence.

Chaume, F. (2004), 'Film studies and translation studies: Two disciplines at stake in audiovisual translation', *Meta* **49**(1), pp. 12–24.

Chen, W. (2006), Explication Through the Use of Connectives in Translated Chinese: A Corpus-based Study, PhD thesis, University of Manchester.

Cheong, H. (2006), 'Target Text Contraction in English-into-Korean Translations: A contradiction of presumed translation universals?', *Meta* **51**(2), pp. 343–367.

Chesterman, A. (2004*a*), *Claims, Changes and Challenges in Translation Studies: Selected Contributions from the EST Congress, Copenhagen 2001*, Amsterdam: John Benjamins Publishing Company, chapter Hypotheses About Translation Universals, pp. 1–13.

Chesterman, A. (2004*b*), *Translation Universals: Do They Exist?*, Amsterdam: John Benjamins Publishing Company, chapter Beyond the Particular, pp. 33–49.

Chesterman, A. (2011), *Handbook of Translation Studies: v. 2*, Amsterdam / Philadelphia: John Benjamins Publishing Company, chapter Translation Universals, pp. 175–179.

Cohen, W. (1995), Fast Effective Rule Induction, In: *Machine Learning: Proceedings of the Twelfth International Conference (ML95)*, Morgan Kaufmann, pp. 115–123.

Corpas, G. (2008), *Investigar con corpus en traducción: los retos de un nuevo paradigma*, Frankfurt am Main, Berlin & New York: Peter Lang.

Corpas, G., Mitkov, R., Afzal, N. and García, L. (2008), Translation universals: do they exist? a corpus-based and nlp approach to convergence, In: *Proceedings of the LREC2008 Workshop on Building and Using Comparable Corpora*, pp. 1–6.

Corpas, G., Mitkov, R., Afzal, N. and Pekar, V. (2008), Translation Universals: Do they exist? A corpus-based NLP study of convergence and simplification, In: *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA 2008)*, Waikiki, Hawaii, pp. 75–81.

Crisafulli, E. (2002), *Crosscultural Transgressions. Research Models in Translation Studies II: Historical and Ideological Issues*, Manchester : St. Jerome Publishing, chapter The Quest for an Eclectic Methodology of Translation Description, pp. 26–43.

Cronin, M. (2003), *Translation and Globalization*, London & New York: Routledge.

De Sutter, G., Delaere, I. and Plevoets, K. (2012), *Quantitative Methods in Corpus-based Translation Studies: a Practical Guide to Descriptive Translation Research*, Amsterdam: John Benjamins Publishing Company, chapter Lexical Lectometry in Corpus-Based Translation Studies. Combining Profile-Based Correspondence Analysis and Logistic Regression Modeling, pp. 325–346.

De Sutter, G. and Van de Velde, M. (2008), Do The Mechanisms That Govern Syntactic Choices Differ Between Original and Translated Language? A Corpus-Based Translation Study of PP Extraposition in Dutch and German, In: R. Xiao, L. He and M. Yue, eds, *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2008)*, Zhejiang University, Hangzhou, China.

De Sutter, G. and Van de Velde, M. (2010), *Using Corpora in Contrastive and Translation Studies*, Newcastle upon Tyne: Cambridge Scholars Publishing, chapter Syntactic Differences Between Translated and Non-translated Dutch: A Corpus-Based In-Depth Analysis of PP Placement, pp. 144–163.

Desmidt, I. (2009), '(Re)translation Revisited', *Meta* **54**(4), pp. 669–683.

Diaz Cintas, J. and Remael, A. (2007), *Audiovisual Translation: Subtitling*, Manchester : St. Jerome Publishing.

Duff, A. (1981), *The Third Language: Recurrent Problems of Translation into English*, Oxford: Pergamon Press.

EAGLES (1996), Expert Advisory Group on Language Engineering Standards Guidelines, Technical report, European Commission, within DG XIII Linguistic Research and Engineering Programme.

Eskola, S. (2002), *Syntetisoivat Rakenteet käännössuornessa*, Joensuu: University of Joensuu.

Fallows, J. (2011), An 'economic burden' Google can no longer bear?, In: *The Atlantic*. [online] Available at: <http://www.theatlantic.com/technology/archive/2011/06/an-economic-burden-google-can-no-longer-bear/240283/>[Accessed 18 August 2012].

Francis, W. N. (1992), Language Corpora B. C., In: J. Svartvik, ed., *Directions in Corpus Linguistics. Trends in Linguistics. Studies and Mongraphs[TILSM]. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, Berlin / New York: Mouton de Gruyter, pp. 17–34.

Frankenberg-Garcia, A. (2004), Are translations longer than source texts? A corpus-based study of explicitation, In: *Third International Corpus Use and Learning to Translate Conference, Barcelona, Spain*, pp. 1–8.

Frawley, W. (1984), *Translation: Literary, Linguistic and Philosophical Perspectives*, Newark: University of Delaware Press, chapter Prolegomenon to a Theory of Translation, pp. 159–175.

Gaspari, F. and Bernardini, S. (2010), *Using Corpora in Contrastive and Translation Studies*, Newcastle upon Tyne: Cambridge Scholars Publishing, chapter Comparing Non-native and Translated Language: Monolingual Comparable Corpora with a Twist, pp. 215–234.

Gellerstam, M. (1986), *Translation Studies in Scandinavia*, Lund: Lund University Press, chapter Translationese in Swedish Novels Translated from English, pp. 88–95.

Gellerstam, M. (1996), Translations as a Source for Cross-linguistic Studies, In: B. A. K. Aijmer and M. Johansson, eds, *Languages in Contrast*, Lund: CWK Gleerup, pp. 53–62.

Gentzler, E. (1993), *Contemporary Translation Theory*, London & New York: Routledge.

Grange, S. (1996), *Language in Contrast: Papers from a Symposium on Text-based Cross-Linguistic Studies*, Lund: Lund University Press., chapter From CA to CIA and Back: an Integrated Approach to Computerised Bilingual and Learner Corpora, pp. 38–51.

Grosz, B. and Sidner, C. (1986), 'Attention, Intention, and the Structure of Discourse', *Computational Linguistics* **12**, pp. 175–204.

Gumul, E. (2004), Cohesion in Interpreting, PhD thesis, Katowice: University of Silesia.

Gumul, E. (2006), 'Explicitation in Simultaneous Interpreting: A Strategy or a By-Product of Language Mediation?', *Across Languages and Cultures* **7**(2), pp. 171–190.

Halliday, M. and Hasan, R. (1976), *Cohesion in English*, London: Longman.

Halverson, S. (1998), 'Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study', *Meta* **43**(4), pp. 494–514.

Hansen, S. (2003), The Nature of Translated Text. An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations, PhD thesis, Saarbrücken: Saarland University.

Harley, T. (2008), *The Psychology of the Language: from Data to Theory*, New York: Psychology Press.

Hatim, B. and Munday, J. (2004), *Translation. An Advanced Resource Book*, London: Routledge.

Helmantel, M. (2002), Interactions in the Dutch adpositional domain, PhD thesis, University of Leiden.

Holmes, J. S. (1988), *Translated!:Papers on Literary Translation and Translation Studies*, Amsterdam: Editions Rodopi.

House, J. (2004), *Neue Perspektiven in der Übersetzungs und Dolmetschwissenschaft*, Bochum: AKS, chapter Explicitness in Discourse Across Languages, pp. 185–208.

Hunston, S. (2002), *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.

Jakobson, R. (1959/2000), *The Translation Studies Reader*, London and New York: Routledge, chapter On Linguistic Aspects of Translation, pp. 113–118.

Jantunen, J. (2001), 'Synonymity and Lexical Simplification in Translations: A Corpus-based Approach', *Across Languages and Cultures* **2**(1), pp. 97–112.

Jantunen, J. (2004*a*), *Incorporating Corpora. The Linguist and the Translator*, Clevedon: Multilingual Matters, chapter Untypical Patterns in Translations. Issues on Corpus Methodology and Synonymity, pp. 101–126.

Jantunen, J. (2004*b*), Synonymity and Translated Finnish. A Corpus-based View of Contextuality of Synonymous Expressions and Lexical Features Specific to Translated Language, PhD thesis, Savonlinna School of Translation Studies, University of Joensuu, Finland.

Johansson, S. (1998), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, Amsterdam: Rodopi, chapter On the Role of Corpora in Cross-linguistic Research, pp. 3–25.

John, G. and Langley, P. (1995), *Estimating Continuous Distributions in Bayesian Classifiers*, Morgan Kaufmann.

Kamenická, R. (2007), 'Defining Explicitation in Translation', *Brno Studies in English* **33**, pp. 45–57.

Kennedy, G. (1998), *An Introduction to Corpus Linguistics*, London: Longman.

Kenny, D. (1998), 'Creatures of a Habit? What Translators Usually Do with Words', *Meta* **43**(4), pp. 515–523.

Kilgarriff, A. (2001), 'Comparing Corpora', *International Journal of Corpus Linguistics* **6**(1), pp. 1–37.

Klaudy, K. (1996), *Translation Studies in Hungary*, Budapest: Scholastica, chapter Back-translation as a Tool for Detecting Explicitation Strategies in Translation, pp. 99–114.

Klaudy, K. (2008), *Routledge Encyclopedia of Translation Studies*, London & New York: Routledge, 2nd edition, chapter Explicitation, pp. 104–108.

Klaudy, K. and Karoli, K. (2005), 'Implicitation in Translation: Empirical Evidence for Operational Asymmetry in Translation', *Across Languages and Cultures* **6**(1), pp. 13–28.

Koehn, P. (2005), Europarl: A parallel Corpus for Statistical Machine Translation, In: *MT Summit, 2005*.

Koppel, M. and Ordan, N. (2011), Translationese and Its Dialects, In: *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), Portland, OR, USA*, pp. 1318–1326.

Kujamäki, P. (2006), *Sociocultural Aspects of Translation and Interpreting*, Amsterdam: John Benjamins Publishing Company, chapter 'Of Course Germans have a Certain Interest in Finland, But...'. Openess to Finnish Literature in Germany in the 1920s and 1930s, pp. 41–52.

Kuncheva, L. I. (2004), *Combining Pattern Classifiers: Methods and Algorithms*, Wiley-Interscience: A John Wiley & Sons, Inc., Publication.

Kurokawa, D., Goutte, C. and Isabelle, P. (2009), Automatic Detection of Translated Text and Its Impact on Machine Translation, In: *Proceedings of the Machine Translation Summit XII, Ottawa, Ontario, Canada*.

Laviosa-Braithwaite, S. (1996), The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation, PhD thesis, University of Manchester.

Laviosa-Braithwaite, S. (1997), *Transferre necesse est. Proceedings of the Second International Conference on Current Trends in Studies of Translation and Interpretating*, Budapest: Scholastica, chapter Investigating Simplification in an English Comparable Corpus of Newspaper Articles, pp. 531–540.

Laviosa, S. (1995), Comparable corpora: Towards a corpus linguistic methodology for the empirical study of translation, In: M. Thelen and

B. Lewandowska-Tomaszczyck, eds, *Translation and Meaning Part III, Second International Maastricht-Lodz Duo Colloquium on 'Translation and Meaning'*, Maastricht: Hogeschool Maastricht.

Laviosa, S. (1998), 'Core Patterns of Lexical Use in Comparable Corpus of English Narrative Prose', *Meta* **43**(4), pp. 557–570.

Laviosa, S. (2002), *Corpus-based Translation Studies. Theory, Findings, Applications*, Amsterdam: Rodopi.

Leech, G. (1992), Corpora and Theories of Linguistic Performance, In: J. Svartvik, ed., *Directions in Corpus Linguistics. Trends in Linguistics. Studies and Mongraphs[TILSM]. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, Berlin / New York: Mouton de Gruyter, pp. 105–122.

Lembersky, G., Ordan, N. and Wintner, S. (2011), Language Models for Machine Translation: Original vs. Translated Texts, In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh: Association for Computational Linguistics, pp. 363–374.

Lembersky, G., Ordan, N. and Wintner, S. (2012), Adapting Translation Models to Translationese Improves SMT, In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics 2012, Avignon, France*, Association for Computational Linguistics, pp. 255–265.

Maia, B. (2003), What are comparable corpora?, In: *Proceedings of the Workshop on Multilingual Corpora: Linguistic Requirements and Technical perspectives, Corpus Linguistics*, Lancaster, United Kingdom, pp. 27–34.

Malmkjaer, K. (1997), *Nonverbal Communication and Translation: New Perspectives and Challenges in Literature, Interpretation and the Media*, Amsterdam: John Benjamins Publishing Company, chapter Punctuation in Hans Christian Andersen's stories and their translations into English, pp. 151–162.

Malmkjaer, K. (2008), *Incorporating Corpora: the Linguist and the Translator*, Clevedon: Multilingual Matters, chapter Norms and Nature in Translation Studies, pp. 49–59.

Marşic, G. (2011), Temporal Processing of News: Annotation of Temporal Expressions, Verbal Events and Temporal Relations, PhD thesis, University of Wolverhampton, Wolverhampton, UK.

Mauranen, A. (2000), *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*, Manchester: St. Jerome Publishing, chapter Strange Strings in Translated Language: A Study on Corpora, pp. 119–141.

Mauranen, A. (2008), *Incorporating Corpora: the Linguist and the Translator*, Multilingual Matters Ltd., chapter Universal Tendencies in Translation, pp. 32–48.

McEnery, A. (2003), *Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, chapter Corpus linguistics, pp. 448–463.

McEnery, A. and Xiao, R. (2007*a*), *Incorporating Corpora. The Linguist and the Translator*, Clevedon: Multilingual Matters, chapter Parallel and Comparable Corpora: What is Happening?, pp. 18–31.

McEnery, A., Xiao, R. and Tono, Y. (2006), *Corpus-Based Language Studies: An Advanced Resource Book*, Routledge.

McEnery, A. and Xiao, Z. (2002), 'English-chinese translation', *Journal of Languages in Contrast* **2**(2), pp. 211–231.

McEnery, T. and Xiao, R. (2007*b*), *Corpus-Based Perspectives in Linguistics*, Amsterdam: John Benjamins Publishing Company, chapter Parallel and Comparable Corpora: The State of Play, pp. 131–145.

Meldrum, Y. F. (2009), Contemporary Translationese in Japanese Popular Literature, PhD thesis, University of Alberta.

Mihăilă, C., Ilisei, I. and Inkpen, D. (2011), 'Zero Pronominal Anaphora Resolution for the Romanian Language', *Research Journal on Computer Science and Computer Engineering with Applications "POLIBITS"* **42**.

Mitchell, T. (1997), *Machine Learning*, New York: McGraw-Hill Science.

Mitchell, T. (2006), The Discipline of Machine Learning, Technical Report CMU-ML-06-108, Machine Learning Department, Carnegie Mellon University.

Mitkov, R. (2002), *Anaphora Resolution*, London: Longman.

Mladin, C. I. (2005), 'Procese şi Structuri Sintactice "Marginalizate" în Sintaxa Românească Actuală. Consideraţii Terminologice Din Perspectivă Diacronică Asupra Contragerii - Construcţiilor - Elipsei', *The Annals of Ovidius University Constanţa - Philology* **16**, pp. 219–234.

Munday, J. (2008), *Introducing Translation Studies. Theories and Applications (Second Edition)*, London & New York: Routledge.

Nevalainen, S. (2005), *Käännössuomeksi. Tutkimuksia Suomennosten Kielestä. Tampere Studies in Language, Translation and Culture A1*, Tampere: Tampere University Press, chapter Köyhtyykö kieli käännettäessä? - Mitä Taajuuslistat Kertovat Suomennosten Sanastosta, pp. 139–160.

Nida, E. (2000), *The Translation Studies Reader*, London: Routledge, chapter Principles of Correspondence, pp. 126–140.

Nida, E. A. (1964), *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*, Brill.

Oakes, M. (2012), *Quantitative Methods in Corpus-based Translation Studies: a Practical Guide to Descriptive Translation Research*, Amsterdam: John Benjamins Publishing Company, chapter Describing a Translational Corpus, pp. 115–147.

Olohan, M. (2004), *Introducing Corpora in Translation Studies*, London: Routledge.

Olohan, M. (2007), 'The Status of Scientific Translation', *Journal of Translation Studies* **10**(1), pp. 131–144.

Olohan, M. and Baker, M. (2000), 'Reported 'that' in Translated English: Evidence for Subconscious Processes of Explicitation?', *Across Languages and Culture* **1**(2), pp. 141–158.

Øverås, L. (1998), 'In Search of the Third Code. An Investigation of Norms in Literary Translation', *Meta* **43**(4), pp. 571–590.

Pápai, V. (2004), *Translation Universals. Do They Exist?*, Amsterdam: John Benjamins Publishing Company, chapter Explicitation: A Universal of Translated Texts?, pp. 143–164.

Pasch, R., Braube, U., Breindl, E. and Wabner, U. (2003), *Handbuch der Deutschen Konnektoren*, Berlin: Mouton de Gruyter.

Perego, E. (2003), 'Evidence of Explicitation in Subtitling: Towards a Categorisation', *Across Languages and Cultures* **4**(1), pp. 63–88.

Platt, J. C. (1999), *Advances in Kernel Methods - Support Vector Learning*, MIT Press, chapter Fast Training of Support Vector Machines using Sequential Minimal Optimization, pp. 185–208.

Puurtinen, T. (2003*a*), 'Explicitating and Implicitating Source Text Ideology', *Across Languages and Cultures* **4**(1), pp. 53–62.

Puurtinen, T. (2003*b*), 'Genre-specific Features of Translationese? Linguistic Differences between Translated and Non-translated Finnish Childrens Literature', *Literary and Linguistic Computing* **18**(4), pp. 389–406.

Puurtinen, T. (2004), *Translation Universals. Do They Exist?*, Amsterdam: John Benjamins Publishing Company, chapter Explicitation of Clausal Relations: A Corpus-based Analysis of Clause Connectives in Translated and Non-translated Finnish Children's Literature, pp. 165–176.

Pym, A. (2004), *The Moving Text: Localization, Translation, and Distribution*, Amsterdam: John Benjamins Publishing Company.

Pym, A. (2005), *New Trends in Translation Studies. In Honour of Kinga Klaudy*, Budapest: Akadmia Kiad, chapter Explaining Explicitation, pp. 29–34.

Pym, A. (2008), *Beyond Descriptive Translation Studies*, Amsterdam: John Benjamins Publishing Company, chapter On Toury's Laws of How Translators Translate, pp. 311–328.

Pym, A. (2011), *Translation Research Projects 3*, Tarragona: Intercultural Studies Group, Universitat Rovira i Virgili, chapter Translation Research Terms: a Tentative Glossary for Moments of Perplexity and Dispute, pp. 75–110.

Quinlan, J. R. (1993), *C4.5: programs for machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Reiss, K. and Vermeer, H. J. (1984), *Grundlegung Einer Allgemeinen Translationstheorie*, Tübingen: Niemeyer.

Resnik, P. and Smith, N. (2003), 'The Web as a Parallel Corpus', *Computational Linguistics* **29**(3), pp. 349–380.

Robin, E. (2010), 'Explicitci a Lektorlt Fordtsokban', *Alkalmazott Nyelvszeti Kzlemnyek, Miskolc, V. vfolyam, 1. Szm* **1**, pp. 179–182.

Rodríguez-Castro, M. (2011), Translationese and Punctuation: An Empirical Study of Translated and Non-translated International Newspaper Articles (English and Spanish), In: *Translation and Interpreting Studies*, Vol. 6, Amsterdam: John Benjamins Publishing Company, chapter 22, pp. 40–61.

Salton, G. and McGill, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York. bag of words model.

Shuttleworth, M. and Cowie, M. (1997), *Dictionary of Translation Studies*, Manchester: St. Jerome Publishing.

Simpson, J. and Weiner, E. (1989), *Oxford English Dictionary Second Edition*, Oxford: Clarendon Press.

Sinclair, J. (1996), *EAGLES Preliminary Recommendations on Corpus Typology. EAGTCWG-CTYP/P*, Pisa: ILC-CNR.

Sinclair, J. (2001), *Small Corpus Studies and ELT Theory and Practice*, Amsterdam: John Benjamins Publishing Company, chapter Preface, pp. vii–xvi.

Stolcke, A. (2002), SRILMan Extensible Language Modelling Toolkit, In: *International Conference on Spoken Language Processing*, pp. 901–904.

Stubbs, M. (1986), *Talking About Text. Studies Presented to David Brazil on His Retirement.*, Discourse Analysis Monograph, Birmingham: English Language Research, University of Birmingham, chapter Lexical Density: A Computational Technique and Some Findings, pp. 27–42.

Taivalkoski, K. (2002), 'Traduire la mixité formelle: L'examples des premières (re)traductions de Fielding en France', *Faits de Langue* **19**, pp. 85–97.

Tapanainen, P. and Järvinen, T. (1997), A Non-projective Dependency Parser, In: *Proceedings of the 5th Conference of Applied Natural Language Processing, Washington D.C., USA*, pp. 64–71.

Teich, E. (2003), *Cross-linguistic Variation in System and Text*, Berlin: Mouton de Gruyter.

Temnikova, I. (2012), Text Complexity and Text Simplification in the Crisis Management Domain, PhD thesis, University of Wolverhampton.

Tirkkonen-Condit, S. (2002), 'Translationese A Myth or an Empirical Fact? A Study Into the Linguistic Identifiability of Translated Language', *Target* **14**(2), pp. 207–220.

Toury, G. (1979), 'Interlanguage and its Manifestations in Translation', *Meta* **24**(2), pp. 223–231.

Toury, G. (1980), *In Search of a Theory of Translation*, Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv.

Toury, G. (1995), *Descriptive Translation Studies and Beyond*, Amsterdam: John Benjamins Publishing Company.

Toury, G. (2004), *Translation Universals: Do They Exist?*, Amsterdam: John Benjamins Publishing Company, chapter Probabilistic Explanations in Translation Studies. Welcome as They Are, Would They Qualify as Universals?, pp. 15–32.

Trosborg, A. (1997), *Text Typology and Translation*, Amsterdam: John Benjamins Publishing Company, chapter Translating Hybrid Political Texts, pp. 145–158.

Tufiş, D., Ştefănescu, D., Ion, R. and Ceauşu, A. (2008), *Advances in Multilingual and Multimodal Information Retrieval (CLEF 2007), Lecture Notes in Computer Science*, Vol. 5152, Springer-Verlag, chapter RACAI's Question Answering System at QA@CLEF 2007, pp. 3284–3291.

Tufiş, D., Ion, R., Ceauşu, A. and Ştefănescu, D. (2008), RACAI's Linguistic Web Services, In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis and D. Tapias, eds, *Proceedings of the Sixth International Conference on Language Resources and Evaluation - LREC 2008*, European Language Ressources Association - ELRA, Marrakech, Morocco, pp. 327–333.

Tyers, F. M. and Alperen, M. (2010), South-East European Times: A parallel corpus of the Balkan languages, In: S. Piperidis, M. Slavcheva and C. Vertan, eds, *Proceedings of the LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages*, European Language Resources Association - ELRA, Valletta, Malta, pp. 49–53.

Tymoczko, M. (1998), 'Computerized Corpora and the Future of Translation Studies', *Meta* **43**(4), pp. 652–659.

Tymoczko, M. (2005), 'Trajectories of Research in Translation Studies', *Meta* **50**, pp. 1082–1097.

van Halteren, H. (2008), Source Language Markers in EUROPARL Translations, In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, University of Manchester, pp. 937–944.

Vanderauwera, R. (1985), *Dutch Novels Translated into English: The Transformation of a "Minority" Literature*, Vol. 6 of *Approaches to Translation Studies*, Amsterdam: Rodopi.

Verdejo, F. M. (1999), The Spanish Wordnet, Technical report, Universitat Politenica de Catalunya, Madrid, Spain.

Vinay, D. (1958), *Stylistique Comparée Du Français Et De l'Anglais*, Didier.

Volansky, V., Ordan, N. and Wintner, S. (2011), More Human or More Translated? Original Texts vs. Human and Machine Translations, In: *11th Bar-Ilan Symposium on the Foundations of Artificial Intelligence (BISFAI 2011)*.

Wang, K. and Qin, H. (2010), *Using Corpora in Contrastive*, Newcastle upon Tyne: Cambridge Scholars Publishing, chapter A Parallel Corpus-based Study of Translational Chinese, pp. 164–181.

Wen, T.-h. (2009), Simplification as a Recurrent Translation Feature: A Corpus-based Study of Modern Chinese Translated Mystery Fiction in Taiwan, PhD thesis, University of Manchester.

Williams, J. and Chesterman, A. (2002), *The Map: A Beginner's Guide to Doing Research in Translation Studies*, Manchester: St. Jerome Publishing.

Witten, I. H., Frank, E. and Hall, M. A. (2011), *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*, Kaufmann, Morgan.

Xiao, R. (2010*a*), Corpus Creation, In: N. Indurkhya and F. J. Damerau, eds, *Handbook of Natural Language Processing*, second edn, Florida: CRC Press, Taylor and Francis Group.

Xiao, R. (2011), 'Word Clusters and Reformulation Markers in Chinese and English: Implications for Translation Universal Hypotheses', *Languages in Contrast* **11**(2), pp. 145–171.

Xiao, R., He, L. and Yue, M. (2010), *Using Corpora in Contrastive and Translation Studies*, Newcastle upon Tyne: Cambridge Scholars Publishing, chapter In Pursuit of the Third Code: Using the ZJU Corpus of Translational Chinese in Translation Studies, pp. 182–214.

Xiao, R. and Yue, M. (2009), *Contemporary Corpus Linguistics*, London: Continuum, chapter Using Corpora in Translation Studies: The State of the Art, pp. 237–262.

Xiao, Z. (2010*b*), 'How Different is Translated Chinese From Native Chinese? A Corpus-based Study of Translation Universals', *International Journal of Corpus Linguistics* **15**(1), p. 333.

Zanettin, F. (2000), *Cross-cultural Transgressions. Research Models in Translation Studies I: Textual and Cognitive Aspects*, Manchester: St. Jerome Publishing, chapter Parallel Corpora in Translation Studies: Issues in Corpus Design, pp. 105–118.